A02

(12) INTERNATIONAL APPLICATION PUBLISHED UNDER THE PATENT COOPERATION TREATY (PCT)

(19) World Intellectual Property Organization
International Bureau

(43) International Publication Date
9 August 2001 (09.08.2001)

PCT

(10) International Publication Number
WO 01/57252 A2

(51) International Patent Classification⁷: C12Q 1/68

(21) International Application Number: PCT/US01/03003

(22) International Filing Date: 29 January 2001 (29.01.2001)

(25) Filing Language: English

(26) Publication Language: English

(30) Priority Data:

60/180,312	4 February 2000 (04.02.2000)	US
60/207,456	26 May 2000 (26.05.2000)	US
09/608,408	30 June 2000 (30.06.2000)	US
09/632,366	3 August 2000 (03.08.2000)	US
60/234,687	21 September 2000 (21.09.2000)	US
60/236,359	27 September 2000 (27.09.2000)	US
0024263.6	4 October 2000 (04.10.2000)	GB

(71) Applicant: **AEOMICA, INC.** [US/US]; 929 East Arques Avenue, Sunnyvale, CA 94086 (US).

(72) Inventors: **PENN, Sharron, Gaynor**; 617 South Delaware Street, San Mateo, CA 94402 (US). **RANK, David, Russell**; 117 El Dorado Commons, Fremont, CA 94539 (US). **HANZEL, David, Kagen**; 988 Loma Verde Avenue, Palo Alto, CA 94303 (US).

(74) Agents: **BECKER, Daniel, M.** et al.; Fish & Neave, 1251 Avenue of the Americas, New York, NY 10020 (US).

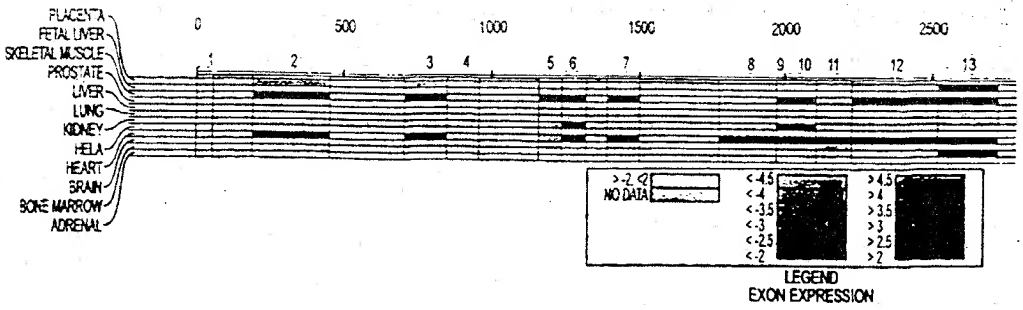
(81) Designated States *(national)*: AE, AG, AL, AM, AT, AU, AZ, BA, BB, BG, BR, BY, BZ, CA, CH, CN, CR, CU, CZ, DE, DK, DM, DZ, EE, ES, FI, GB, GD, GE, GH, GM, HR, HU, ID, IL, IN, IS, JP, KE, KG, KP, KR, KZ, LC, LK, LR, LS, LT, LU, LV, MA, MD, MG, MK, MN, MW, MX, MZ, NO, NZ, PL, PT, RO, RU, SD, SE, SG, SI, SK, SL, TJ, TM, TR, TT, TZ, UA, UG, UZ, VN, YU, ZA, ZW.

(84) Designated States *(regional)*: ARIPO patent (GH, GM, KE, LS, MW, MZ, SD, SL, SZ, TZ, UG, ZW), Eurasian patent (AM, AZ, BY, KG, KZ, MD, RU, TJ, TM), European patent (AT, BE, CH, CY, DE, DK, ES, FI, FR, GB, GR, IE, IT, LU, MC, NL, PT, SE, TR), OAPI patent (BF, BJ, CF, CG, CI, CM, GA, GN, GW, ML, MR, NE, SN, TD, TG).

Published:
without international search report and to be republished upon receipt of that report

For two-letter codes and other abbreviations, refer to the "Guidance Notes on Codes and Abbreviations" appearing at the beginning of each regular issue of the PCT Gazette.

(54) Title: METHODS AND APPARATUS FOR HIGH-THROUGHPUT DETECTION AND CHARACTERIZATION OF ALTERNATIVELY SPLICED GENES



(57) Abstract: Methods and apparatus for designing and producing single exon probes from genomic sequence data are presented. Also presented are genome-derived single exon microarrays. The single exon probes and genome-derived microarrays are used for high-throughput interrogation of exon-specific expression in a plurality of tissues and cell types. Alternative splice events are detected as reproducible changes in relative or absolute expression of exons. Visual tools and automated methods for detecting and characterizing the alternative splice events are presented.

WO 01/57252 A2

METHODS AND APPARATUS FOR HIGH-THROUGHPUT DETECTION AND CHARACTERIZATION OF ALTERNATIVELY SPLICED GENES

FIELD OF THE INVENTION

The present invention is in the field of molecular biology, and particularly relates to methods and apparatus for high throughput identification and characterization of alternatively spliced eukaryotic genes, and to bioinformatic methods relating thereto.

BACKGROUND OF THE INVENTION

Alternative splicing, first predicted in 1978, Gilbert, *Nature* 271:501 (1978), was soon thereafter demonstrated to provide an important supplement to transcriptional control in a variety of eukaryotic organisms, both in the regulation of cell type-specific expression and in the regulation of developmental stage-specific expression of structurally distinct proteins. By 1987 more than 50 genes, in organisms ranging from *Drosophila* to humans, had been shown to generate protein diversity through use of

alternative splicing. Breitbart *et al.*, *Ann. Rev. Biochem.* 56:467-95 (1987).

Recent reports suggest that at least one-third, and likely a higher percentage, of human genes are alternatively spliced. Hanke *et al.*, *Trends Genet.* 15(1):389 - 390 (1999); Mironov *et al.*, *Genome Res.* 9:1288-93 (1999); Brett *et al.*, *FEBS Lett.* 474(1):83-6 (2000). With estimates of the number of human genes ranging from 35,000, Ewing *et al.*, *Nature Genet.* 25(2):232-234 (2000), to 120,000, Liang *et al.*, *Nature Genet.* 25(2):239-240 (2000), and with the *Drosophila* homolog of one human gene reported to have 38,000 potential alternatively spliced variants, Schmucker *et al.*, *Cell* 101:671 (2000), it now appears that alternative splicing, long recognized as important in specialized cases, is not only ubiquitous, but indeed may permit the human genome to encode millions, perhaps tens of millions, of structurally distinct proteins and protein isoforms.

Historically, identification and characterization of splice variants has proceeded on an essentially *ad hoc* basis. Identification has, e.g., been made through the fortuitous observation of multiple sizes of an mRNA on Northern blot, of disparities in restriction map or of nucleic acid sequence of multiple cDNA clones obtained from a single gene, of differences in structure and sequence of RT-PCR products, and through observed homology to genes known to be alternatively spliced in other species. Further characterization of the presumed alternative splicing events has typically proceeded by comparison of cDNA to genomic sequence, analysis of genomic

sequence to identify consensus splice acceptor and donor sequences, and through use of fragmentary probes to assess the tissue-specific expression of the identified variants. *See, e.g.*, Nakamura *et al.*, *J. Biol. Chem.* 270(50):30102-30110 (1995); Kramer *et al.*, *Gene* 211(1):29-37 (1998); Screaton *et al.*, *Proc. Natl. Acad. Sci. USA* 89:12160-12164 (1992).

With the increased availability of both genomic and message-based sequence data, the latter typically in the form of partial cDNAs or ESTs, more recent efforts to identify alternatively spliced variants have been based upon bioinformatic approaches.

Hanke *et al.*, *Trends Genet.* 15(10):389-390 (1999) searched for alternatively spliced human genes by aligning six frame translations of about 1.3 million human ESTs to the amino acid sequences of 475 human disease-associated proteins. A difference in length of alignment match between the EST and protein sequence was used as the criterion for calling potential alternatively spliced forms. Using this method, the authors predicted 222 candidate alternative splice sites in 162 of the 475 proteins (34% of proteins).

Mironov *et al.*, *Genome Res.* 9:1288-93 (1999) searched for potential alternatively spliced genes by aligning EST contigs to genomic DNA. The existence of a plurality of such assembled EST superstructures for a gene was used as the criterion for calling a potential alternative splicing event. Using this method, the authors suggest that more than one-third of human genes have at least two variants of exon-intron structure.

Croft *et al.*, *Nature Genet.* 24:340-341 (2000) identified potential alternative splicing events by

querying EST databases with sequences previously annotated as being intronic, thus identifying coding regions that had been spliced out of known messages.

Each of the aforementioned approaches to identifying alternatively spliced genes relies upon a prior determination of cDNA (or, equivalently, of EST) sequence, whether directly or through use of sequence data prior-accessioned into a database. This reliance renders these methods poorly suited to the identification, and particularly to the high throughput identification, of the large number of splice variants now expected to be expressed by eukaryotic organisms.

For example, we have recently demonstrated that the genes represented in EST and other expression databases represent only a fraction of expressed genes: no more than about 1/3 of expressed human genes (in any of the gene's potential alternatively spliced forms) are represented in existing expression databases, notwithstanding the fact that the human genome has the greatest depth of EST coverage of any eukaryotic organism. *See* commonly owned and co-pending provisional U.S. patent application serial nos. 60/180,312, filed February 4, 2000; 60/207,456, filed May 26, 2000; 60/234,687, filed September 21, 2000; 60/236,359, filed September 27, 2000; and U.K. patent application no. 24263.6, filed October 4, 2000, incorporated herein by reference in their entireties. If splice variants are uniformly distributed among the genes present in and absent from EST databases, fully 2/3 of genes with potential splice variants are at present inaccessible to EST/cDNA-reliant identification methods.

Hanke *et al.*, *supra*, further note that EST libraries provide only fragmentary coverage for even that subset of genes represented therein: ESTs that matched the 475 template proteins covered only about 50% of each protein sequence, making impossible predictions about alternative splice sites in the remainder of the proteins.

Representation of genes in EST libraries is not only incomplete and fragmentary, but strongly biased as well; methods for identifying splice variants that rely upon such data will share such bias.

It is, for example, axiomatic that such libraries are biased by the tissue or cell type of message origin. Mironov *et al.*, *supra*, point out that many alternatively spliced variants will have very limited cell-type or stage specificity, and that such variants will, accordingly, be particularly under-represented in available EST and cDNA databases, rendering them poorly detectable by present methods. We have recently demonstrated that representation in EST and expression databases is further and significantly biased toward genes with higher expression levels. *See* U.S. Provisional patent applications nos. 60/180,312, filed February 4, 2000 and 60/207,456, filed May 26, 2000, incorporated herein by reference in their entireties. These data imply that splice variants expressed at low levels will be under-represented in available EST and cDNA databases; like variants with limited cell-type or stage specificity, such low expression variants will prove difficult to detect with present methods.

Reliance upon EST and cDNA sequence imposes other sources of error. Intron retention in the EST

libraries (and thus in corresponding sequence databases), through either genomic contamination, incomplete/incorrect splicing, or the presence of unprocessed pseudogenes, will lead to an overestimate of alternative splice events by bioinformatic approaches. Mironov *et al.*, *supra*; Hanke *et al.*, *supra*. Mironov *et al.* further comment that their particular algorithm is capable of generating artifactual chimeric EST contigs – the bioinformatic equivalent of cloning artifacts known to populate EST and cDNA libraries – leading to an overestimate of the number of alternative splice forms per gene.

There thus exists a need in the art for methods of identifying alternatively spliced variants of eukaryotic genes that do not rely upon the prior cloning and/or sequencing of message. Given the potentially vast number of such alternatively spliced variants, there is a particular need for methods that are suitable for high throughput identification of splice variants.

Present bioinformatic methods for identifying alternatively spliced genes are capable only of identifying potential variants. Confirmation of the existence of each of the proposed variants, and a full understanding of the regulatory and developmental gene expression programs effected through alternative splicing, requires that the proposed variants be correlated experimentally with patterns of expression.

There thus exists a need for methods and apparatus that can be used to measure the expression of alternatively spliced gene variants in a plurality of separate tissues and cells at discrete developmental stages and under various physiologic conditions. Croft

et al., supra, suggest that microarray systems for analyzing gene expression will need to be expanded beyond a single coding sequence per gene to include all possible alternative exons on the array, in order to understand the genetic output and cellular circuitry of higher organisms; Croft *et al.* suggest no means, however, to identify all possible alternative exons of the gene to be assayed.

Given the large number of variants expected in higher eukaryotes, there is a need for automated methods for identifying alternative splice variants from large quantities of expression data.

The SWISSPROT database provides special feature lines for annotated splice variants that have been previously reported in the literature. The ASDB database, Gelfand *et al., Nucl. Acids Res.* 27(1):301-302 (1999), contains entries extracted from SWISSPROT in which the SWISSPROT annotation contained the words "alternative splicing", further clustered in the ASDB database according to presence of common fragments not shorter than 20 amino acids. Given the large number of alternatively spliced variants predicted in the human and other eukaryotic organisms, and the large number of cell types, developmental stages, and physiologic conditions that must be assayed for each, there is a need in the art for bioinformatic methods and apparatus for meaningfully relating the expression information of alternatively spliced variants to the sequence data. There is an accompanying need for methods and apparatus for meaningfully displaying such information.

SUMMARY OF THE INVENTION

In view of the foregoing, it is an object of the present invention to provide methods for identifying alternatively spliced variants of eukaryotic genes that do not rely upon the prior cloning and/or sequencing of transcribed message. It is a particular object to provide methods that are suitable for high throughput use.

It is a further object of the present invention to provide methods and apparatus that can be used to measure the expression of alternatively spliced gene variants in a plurality of separate tissues and cells at discrete developmental stages and under various physiologic conditions.

It is a still further object of the invention to provide methods for identifying alternative splice variants from large quantities of expression data.

It is yet another object of the invention to provide bioinformatic methods and apparatus for meaningfully relating the expression information of alternatively spliced variants to the sequence data, and to provide methods and apparatus for meaningfully displaying such information.

These and other objects of the present invention are achieved by providing methods and apparatus for generating single exon probes from genomic sequence data. The single exon probes, each typically comprising an exon with flanking intergenic or intronic sequence, can be used to interrogate exon-specific expression in a plurality of tissues or cell types. The invention further provides genome-derived, single exon microarrays that facilitate the high

throughput interrogation of exon-specific expression in a plurality of tissues or cell types. The exon-specific expression data permit a variety of alternative splice events to be detected and characterized. The invention provides manual methods for such detection and characterization, visual display tools particularly adapted to such detection, and automated methods implemented on a computer.

BRIEF DESCRIPTION OF THE DRAWINGS

The above and other objects and advantages of the present invention will be apparent upon consideration of the following detailed description taken in conjunction with the accompanying drawings, in which like characters refer to like parts throughout, and in which:

FIG. 1 illustrates a process for predicting functional regions from genomic sequence, confirming the functional activity of such regions experimentally, and associating and displaying the data so obtained in meaningful and useful relationship to the original sequence data, according to the present invention;

FIG. 2 further elaborates that portion of the process schematized in FIG. 1 for predicting functional regions from genomic sequence, according to the present invention;

FIG. 3 illustrates a visual display according to the present invention, herein denominated a "Mondrian", in which a single genomic sequence is annotated with predicted and experimentally confirmed functional information;

FIG. 4 presents a Mondrian of a hypothetical annotated genomic sequence, further identifying typical color conventions when the Mondrian is used to annotate genomic sequence with exon-specific expression data;

FIG. 5 presents a Mondrian of BAC AC008172 (bases 25,000 to 130,000), containing the carbamyl phosphate synthetase gene (AF154830.1);

FIG. 6 presents a Mondrian of BAC A049839;

FIG. 7 is a chart that summarizes data from experimental Example 1, showing the size distributions of predicted exon length (dashed line) and actual PCR products (amplicons) (solid line) as obtained from human genomic sequence according to the methods of the present invention;

FIG. 8 is a histogram that summarizes data from experimental Examples 1 and 2, showing the number of tissues in which predicted exons could be shown to be expressed using simultaneous two color hybridization to a genome-derived single exon microarray of the present invention. The graph shows the number of sequence-verified products that were either not expressed in any of the ten tested tissues/cell types ("0"), expressed in one or more but not all tested tissues ("1" - "9"), or expressed in all tissues tested ("10");

FIG. 9 is a pictorial representation of data from experimental Examples 1 and 2, showing the expression (ratio relative to control) of probes having verified sequences that were expressed with signal intensity greater than 3 in at least one tissue, with: FIG. 9A showing both the expression as measured by microarray hybridization in each of the 10 measured tissues and the expression as measured

"bioinformatically" by query of EST, NR and SwissProt databases; with FIG. 9B showing the legend for display of physical expression (ratio) in FIG. 9A; and with FIG. 9C showing the legend for scoring EST hits as depicted in FIG. 9A;

FIG. 10 is a chart of data from experimental Examples 1 and 2, showing a comparison of normalized CY3 signal intensity for arrayed sequences that were identical to sequences in existing EST, NR and SwissProt databases (known) or that were dissimilar (unknown), where the dashed line denotes the signal intensity for all sequence-verified products with a BLAST Expect ("E") value of greater than 1e-30 (1×10^{-30}) ("unknown") and the solid line denotes sequence-verified spots with a BLAST expect ("E") value of less than 1e-30 (1×10^{-30}) ("known");

FIGS. 11A and 11B show two views of a genome-derived single-exon microarray particularly designed to detect alternative splice events, after simultaneous two-color hybridization to human kidney cDNA and control cDNA, with FIG. 11A showing an enlarged partial view of one field of the slide shown in full in FIG. 11B;

FIG. 12 shows a modified Mondrian "splice viewer" of a known gene for which high throughput exon-specific expression analysis with a genome-derived single exon microarray demonstrates alternative splicing; and

FIG. 13 shows a modified Mondrian "splice viewer" of an unknown gene for which high throughput exon-specific expression analysis with a genome-derived single exon microarray demonstrates alternative splicing.

DETAILED DESCRIPTION OF THE INVENTION

Definitions

As used herein, the phrase "alternative splicing" and its linguistic equivalents includes all types of RNA processing that lead to expression of plural protein isoforms from a single gene; accordingly, the phrase "splice variant(s)" and its linguistic equivalents embraces mRNAs transcribed from a given gene that, however processed, collectively encode plural protein isoforms.

For example, and by way of illustration only, splice variants can include exon insertions, exon extensions, exon truncations, exon deletions, alternatives in the 5' untranslated region ("5' UT") and alternatives in the 3' untranslated region ("3' UT"). Such 3' alternatives include, for example, differences in the site of RNA transcript cleavage and site of poly(A) addition. See, *e.g.*, Gautheret *et al.*, *Genome Res.* 8:524-530 (1998).

As used herein, the term "microarray" and equivalent phrase "nucleic acid microarray" refer to a substrate-bound collection of plural nucleic acids, hybridization to each of the plurality of bound nucleic acids being separately detectable. The substrate can be solid or porous, planar or non-planar, unitary or distributed.

As so defined, the term "microarray" and phrase "nucleic acid microarray" include all the devices so called in Schena (ed.), DNA Microarrays: A Practical Approach (Practical Approach Series), Oxford University Press (1999) (ISBN: 0199637768); *Nature*

Genet. 21(1)(suppl):1 - 60 (1999); and Schena (ed.), Microarray Biochip: Tools and Technology, Eaton Publishing Company/BioTechniques Books Division (2000) (ISBN: 1881299376), the disclosures of which are incorporated herein by reference in their entireties.

As so defined, the term "microarray" and phrase "nucleic acid microarray" also include substrate-bound collections of plural nucleic acids in which the nucleic acids are distributably disposed on a plurality of beads, rather than on a unitary planar substrate, as is described, *inter alia*, in Brenner *et al.*, *Proc. Natl. Acad. Sci. USA* 97(4):166501670 (2000), the disclosure of which is incorporated herein by reference in its entirety; in such case, the term "microarray" and phrase "nucleic acid microarray" refer to the plurality of beads in aggregate.

As used herein with respect to a nucleic acid microarray, the term "probe" refers to the nucleic acid that is, or is intended to be, bound to the substrate. As used herein with respect to solution phase hybridization, the term "probe" refers to the nucleic acid of known sequence that is, or is intended to be, detectably labeled. In either such context, the term "target" refers to nucleic acid intended to be bound to probe by Watson-Crick complementarity.

As used herein, the expression "probe comprising SEQ ID NO", and variants thereof, intends a nucleic acid probe, at least a portion of which probe has either (i) the sequence directly as given in the referenced SEQ ID NO, or (ii) a sequence complementary to the sequence as given in the referenced SEQ ID NO, the choice as between sequence directly as given and

complement thereof dictated by the requirement that the probe be complementary to the desired target.

As used herein, the phrase "expression of a probe" and its linguistic variants means that the probe hybridizes detectably at high stringency to nucleic acids that derive from mRNA.

As used herein, the term "exon" refers to a nucleic acid sequence bioinformatically predicted to encode a portion of a natural protein.

As used herein, the phrase "open reading frame" and the equivalent acronym "ORF" refer to that portion of an exon that can be translated in its entirety into a sequence of contiguous amino acids. As so defined, an ORF is wholly contained within its respective exon and has length, measured in nucleotides, exactly divisible by 3. As so defined, an ORF need not encode the entirety of a natural protein.

As used herein, the phrase "specific binding pair" intends a pair of molecules that bind to one another with high specificity. Binding pairs typically have affinity or avidity of at least 10^7, preferably at least 10^8, more preferably at least 10^9 liters/mole. Nonlimiting examples of specific binding pairs are: antibody and antigen; biotin and avidin; and biotin and streptavidin.

As used herein with respect to the visual display of annotated genomic sequence, the term "rectangle" means any geometric shape that has at least a first and a second border, wherein each of the first and second borders is capable of mapping uniquely to a point of another visual object of the display.

Genome-Derived Single Exon Probes

In a first aspect, the present invention presents methods and apparatus for generating single exon probes from genomic sequence data. The single exon probes, each typically comprising an exon with flanking intergenic or intronic sequence, can be used for the high throughput interrogation of exon-specific expression, which in turn permits a variety of alternative splice events to be detected and characterized.

FIG. 1 is a flow chart illustrating in broad outline a process of the present invention for predicting functional regions from genomic sequence, confirming and characterizing the functional activity of such regions experimentally, and then associating and displaying the information so obtained in meaningful and useful relationship to the original genomic sequence data. To generate exon-specific probes for use in alternative splice detection, the functional activity to be predicted, confirmed, associated and displayed is the ability of the predicted region to code for protein.

The initial input into process 10 of the present invention is drawn from one or more databases 100 containing genomic sequence data. Because genomic sequence is usually obtained from subgenomic fragments, the sequence data typically will be stored in a series of records corresponding to these subgenomic sequenced fragments. Some fragments will have been catenated to form larger contiguous sequences ("contigs"); others will not. A finite percentage of sequence data in the database will typically be

erroneous, consisting *inter alia* of vector sequence, sequence created from aberrant cloning events, sequence of artificial polylinkers, and sequence that was erroneously read.

Each sequence record in database 100 will minimally contain as annotation a unique sequence identifier (accession number), and will typically be annotated further to identify the date of accession, species of origin, and depositor. Because database 100 can contain nongenomic sequence, each sequence will typically be annotated further to permit query for genomic sequence. Chromosomal origin, optionally with map location, can also be present. Data can be, and over time increasingly will be, further annotated with additional information, in part through use of the present invention, as described below. Annotation can be present within the data records, in information external to database 100 and linked to the records thereto, or through a combination of the two.

Databases useful as genomic sequence database 100 in the present invention include GenBank, and particularly include several divisions thereof, including the htgs (draft), NT (nucleotide, command line), and NR (nonredundant) divisions. GenBank is produced by the National Institutes of Health and is maintained by the National Center for Biotechnology Information (NCBI). Databases of genomic sequence from species other than human, such as mouse, rat, *Arabidopsis thaliana*, *C. elegans*, *C. brigsii*, *Drosophila melanogaster*, zebra fish, and other higher eukaryotic organisms will also prove useful as genomic sequence database 100.

Genomic sequence obtained by query of genomic sequence database 100 is then input into one or more processes 200 for identification of regions therein that are predicted to have a biological function as specified by the user. Such functions include, but are not limited to, encoding protein, regulating transcription, regulating message transport after transcription, regulating message splicing after transcription, regulating message degradation after transcription, contributing to or controlling chromosomal somatic recombination, contributing to chromosomal stability or movement, contributing to allelic exclusion or X chromosome inactivation, and the like. To generate exon-specific probes for detection of alternative splicing, the functional activity to be predicted, confirmed, associated and displayed will be protein coding.

The particular genomic sequence to be input into process 200 will depend upon the function for which relevant sequence is to be identified as well as upon the approach chosen for such identification. Process step 200 can be iterated to identify different functions within a given genomic region. In such case, the input often will be different for the several iterations.

Sequences predicted to have the requisite function by process 200 are then input into process 300, where a subset of the input sequences suitable for experimental confirmation is identified. Experimental confirmation can involve physical and/or bioinformatic assay. Where the subsequent experimental assay is bioinformatic, rather than physical, there are fewer constraints on the sequences that can be tested,

and in this latter case therefore process 300 can output the entirety of the input sequence.

The subset of sequences output from process 300 is then used in process 400 for experimental verification and characterization of the function predicted in process 200, which experimental verification can, and often will, include both physical and bioinformatic assay.

Process 500 annotates the sequence data with the functional information obtained in the physical and/or bioinformatic assays of process 400. Such annotation can be done using any technique that usefully relates the functional information to the sequence, as, for example, by incorporating the functional data into the sequence data record itself, by linking records in a hierarchical or relational database, by linking to external databases, by a combination thereof, or by other means well known within the database arts. The data can even be submitted for incorporation into databases maintained by others, such as GenBank, which is maintained by NCBI.

As further noted in FIG. 1, additional annotation can be input into process 500 from external sources 600.

The annotated data is then optionally displayed in process 800, either before, concomitantly with, or after optional storage 700 on nontransient media, such as magnetic disk, optical disc, magnetooptical disk, flash memory, or the like.

FIG. 1 shows that the experimental data output from process 400 can be used in each preceding step of process 10: ***e.g.*, facilitating identification**

of functional sequences in process 200, facilitating identification of an experimentally suitable subset thereof in process 300, and facilitating creation of physical and/or informational substrates for, and performance of subsequent assay, of functional sequences in process 400.

Information from each step can be passed directly to the succeeding process, or stored in permanent or interim form prior to passage to the succeeding process. Often, data will be stored after each, or at least a plurality, of such process steps. Any or all process steps can be automated.

FIG. 2 further elaborates the prediction of functional sequence within genomic sequence according to process 200.

Genomic sequence database 100 is first queried 20 for genomic sequence.

The sequence required to be returned by query 20 will depend, in the first instance, upon the function to be identified.

For example, genomic sequences that function to encode protein can be identified *inter alia* using gene prediction approaches, comparative sequence analysis approaches, or combinations of the two. In gene prediction analysis, sequence from one genome is input into process 200 where at least one, preferably a plurality, of algorithmic methods are applied to identify putative coding regions. In comparative sequence analysis, by contrast, corresponding, *e.g.*, syntenic, sequence from a plurality of sources, typically a plurality of species, is input into process 200, where at least one, possibly a plurality, of

algorithmic methods are applied to compare the sequences and identify regions of least variability.

The exact content of query 20 will also depend upon the database queried. For example, if the database contains both genomic and nongenomic sequence, perhaps derived from multiple species, and the function to be predicted is protein coding in human genomic DNA, the query will accordingly require that the sequence returned be genomic and derived from humans.

Query 20 can also incorporate criteria that compel return of sequence that meets operative requirements of the subsequent analytical method. Alternatively, or in addition, such operative criteria can be enforced in subsequent preprocess step 24.

For example, if the function sought to be identified is protein coding, as for generating single exon probes, query 20 can incorporate criteria that return from genomic sequence database 100 only those sequences present within contigs sufficiently long as to have obviated substantial fragmentation of any given exon among a plurality of separate sequence fragments.

Such criteria can, for example, consist of a required minimal individual genomic sequence fragment length, such as 10 kb, more typically 20 kb, 30 kb, 40kb, and preferably 50 kb or more, as well as an optional further or alternative requirement that sequence from any given clone, such as a bacterial artificial chromosome ("BAC"), be presented in no more than a finite maximal number of fragments, such as no more than 20 separate pieces, more typically no more than 15 fragments, even more typically no more than about 10 - 12 fragments.

Results using the present invention have shown that genomic sequence from bacterial artificial chromosomes (BACs) is sufficient for gene prediction analysis according to the present invention if the sequence is at least 50 kb in length, and if additionally the sequence from any given BAC is presented in fewer than 15, and preferably fewer than 10, fragments. Accordingly, query 20 can incorporate a requirement that data accessioned from BAC sequencing be in fewer than 15, preferably fewer than 10, fragments.

An additional criterion that can be incorporated into the query can be the date, or range of dates, of sequence accession. Although the process has been described above as if genomic sequence database 100 were static, it is of course understood that the genomic sequence databases need not be static, and indeed are typically updated on a frequent, even hourly, basis.

Thus, as further described in experimental Examples 1 and 2, *infra*, it is possible to query the database for newly added sequence, either newly added after an absolute date or newly added relative to a prior analysis performed using the methods and apparatus of the present invention. In this way, the process herein described can incorporate a dynamic, temporal component.

One utility of such temporal limitation is to identify, from newly accessioned genomic sequence, the presence of novel genes, particularly those not previously identified by EST sequencing (or other sequencing efforts that are similarly based upon gene expression). As further described in Example 1, such

an approach has shown that newly accessioned human genomic sequence, when analyzed for sequences that function to encode protein, readily identifies genes that are novel over those in existing EST and other expression databases. In fact, as shown below, fully 2/3 of genes identified in newly accessioned human genomic sequence have not hitherto been identified. This makes the methods of the present invention extremely powerful gene discovery tools. And as would be appreciated, such gene discovery can be performed using genomic sequence from species other than human.

If query 20 incorporates multiple criteria, such as above-described, the multiple criteria can be performed as a series of separate queries or as a single query, depending in part upon the query language, the complexity of the query, and other considerations well known in the database arts.

If query 20 returns no genomic sequence meeting the query criteria, the negative result can be reported by process 22, and process 200 (and indeed, entire process 10) ended 23, as shown. Alternatively, or in addition to report and termination of the initial inquiry, a new query 20 can be generated that takes into account the initial negative result.

When query 20 returns sequence meeting the query criteria, the returned sequence is then passed to optional preprocessing 24, suitable and specific for the desired analytical approach and the particular analytical methods thereof to be used in process 25.

Preprocessing 24 can include processes suitable for many approaches and methods thereof, as well as processes specifically suited for the intended subsequent analysis.

Preprocessing 24 suitable for most approaches and methods will include elimination of sequence irrelevant to, or that would interfere with, the subsequent analysis. Such sequence includes repetitive sequence, such as Alu repeats and LINE elements, vector sequence, artificial sequence, such as artificial polylinkers, and the like. Such removal can readily be performed by identification and subsequent masking of the undesired sequence.

Identification can be effected by comparing the genomic sequence returned by query 20 with public or private databases containing known repetitive sequence, vector sequence, artificial sequence, and other artifactual sequence. Such comparison can readily be done using programs well known in the art, such as CROSS_MATCH or REPEATMASKER, the latter available on-line at http://ftp.genome.washington.edu/RM/RepeatMasker.html, or by proprietary sequence comparison programs the engineering of which is well within the skill in the art.

Alternatively, or in addition, undesirable, including artifactual, sequence can be identified algorithmically without comparison to external databases and thereafter removed. For example, synthetic polylinker sequence can be identified by an algorithm that identifies a significantly higher than average density of known restriction sites. As another example, vector sequence can be identified by algorithms that identify nucleotide or codon usage at variance with that of the bulk of the genomic sequence.

Once identified, undesired sequence can be removed. Removal can usefully be done by masking the

undesired sequence as, for example, by converting the specific nucleotide references to one that is unrecognized by the subsequent bioinformatic algorithms, such as "X". Alternatively, but at present less preferred, the undesired sequence can be excised from the returned genomic sequence, leaving gaps.

Preprocessing 24 can further include selection from among duplicative sequences of that one sequence of highest quality. Higher quality can be measured as a lower percentage of, fewest number of, or least densely clustered occurrence of ambiguous nucleotides, defined as those nucleotides that are identified in the genomic sequence using symbols indicating ambiguity. Higher quality can also or alternatively be valued by presence in the longest contig.

Preprocessing 24 can, and often will, also include formatting of the data as specifically appropriate for passage to the analytical algorithms of process 25. Such formatting can and typically will include, *inter alia*, addition of a unique sequence identifier, either derived from the original accession number in genomic sequence database 100, or newly applied, and can further include additional annotation. Formatting can include conversion from one to another sequence listing standard, such as conversion to or from FASTA or the like, depending upon the input expected by the subsequent process.

Preprocessing, which can be optional depending upon the function desired to be identified and the informational requirements of the methods for effecting such identification, is followed by sequence

processing 25, where sequences with the desired function are identified within the genomic sequence.

As mentioned above, such functions can include, but are not limited to, encoding protein, regulating transcription, regulating message transport after transcription, regulating message splicing after transcription, regulating message degradation after transcription, contributing to or controlling chromosomal somatic recombination, contributing to chromosomal stability or movement, contributing to allelic exclusion or X chromosome inactivation, and the like.

Where the function is protein coding, the above-described process of the present invention can be used to identify individual exons from genomic sequence. As described in further detail in commonly owned and copending U.S. provisional application nos. 60/207,456, filed May 26, 2000; 60/234,687, filed September 21, 2000; 60/236,359, filed September 27, 2000; and U.K. patent application no. 24263.6, filed October 4, 2000, the disclosures of which are incorporated herein by reference in their entirety, we have used this process to identify more than 15,000 exons in human genomic sequence whose expression we have confirmed in at least one human tissue or cell type. Fully two-thirds of the exons belong to genes that were not then represented in existing public expression (EST, cDNA) databases.

To identify such individual exons from genomic sequence, process 25 is used to identify putative coding regions. Two exemplary approaches in process 25 for identifying sequence that encodes

putative genes are gene prediction and comparative sequence analysis.

Gene prediction can be performed using any of a number of algorithmic methods, embodied in one or more software programs, that identify open reading frames (ORFs) using a variety of heuristics, such as GRAIL, DICTION, GENSCAN, and GENEFINDER. Comparative sequence analysis similarly can be performed using any of a variety of known programs that identify regions with lower sequence variability.

As further described in Example 1, below, gene finding software programs yield a range of results. For the newly accessioned human genomic sequence input in Example 1, for example, GRAIL identified the greatest percentage of genomic sequence as putative coding region, 2% of the data analyzed; GENEFINDER was second, calling 1%; and DICTION yielded the least putative coding region, with 0.8% of genomic sequence called as coding region.

Increased reliability can be obtained when consensus is required among several such methods. Although discussed herein particularly with respect to exon calling, consensus among methods will in general increase reliability of predicting other functions as well.

Thus, as indicated by query 26, sequence processing 25, optionally with preprocessing 24, can be repeated with a different method, with consensus among such iterations determined and reported in process 27.

Process 27 compares the several outputs for a given input genomic sequence and identifies consensus among the separately reported results. The consensus itself, as well as the sequence meeting that consensus,

is then stored in process 29a, displayed in process 29b, and/or output to process 300 for subsequent identification of a subset thereof suitable for assay.

Multiple levels of consensus can be calculated and reported by process 27.

For example, as further described in Example 1, *infra*, process 27 can report consensus as between all specific pairs of methods of gene prediction, as consensus among any one or more of the pairs of methods of gene prediction, or as among all of the gene prediction algorithms used. Thus, in Example 1, process 27 reported that GRAIL and GENEFINDER programs agreed on 0.7% of genomic sequence, that GRAIL and DICTION agreed on 0.5% of genomic sequence, and that the three programs together agreed on 0.25% of the data analyzed. Put another way, 0.25% of the genomic sequence was identified by all three of the programs as containing putative coding region.

As another example, three of the four gene prediction algorithms that we presently use — GENEFINDER, GENSCAN, and GRAIL — predict frame information in addition to the position of exons. If there is overlap in position and frame of the predicted exons, even if not complete identity, the predicted exons are merged in process 27 to generate the largest possible consensus coding region. The process is iterated until all possible overlaps have been merged. This approach reduces the mean number of exons present in each amplicon, and is preferred in generating exon-specific probes useful for detecting exon elongation and exon truncation alternative splice events, as will be described in more detail below.

Furthermore, consensus can be required among different approaches to identifying a chosen function.

For example, if the function desired to be identified is coding of protein sequence, and a first used approach to exon calling is gene prediction, the process can be repeated on the same input sequence, or subset thereof, with another approach, such as comparative sequence analysis. In such a case, where comparative sequence analysis follows gene prediction, the comparison can be performed not only on genomic nucleic acid sequence, but additionally or alternatively can be performed on the predicted amino acid sequence translated from the exons prior identified by the gene prediction approach.

Although shown as an iterative process, the multiple analyses required to achieve consensus can be done in series, in parallel, or some combination thereof.

Predicted functional sequence, optionally representing a consensus among a plurality of methods and approaches for determination thereof, is passed to process 300 for identification of a subset thereof for functional assay.

Where the function sought to be identified is protein coding, process 300 is used to identify a subset thereof suitable for experimental verification by physical and/or bioinformatic approaches.

Where the goal is the identification and confirmation of expression of only a single exon of gene — for example, to provide a gene-specific probe — putative exons identified in process 200 can be classified, or binned, bioinformatically into putative genes. This binning can be based *inter alia* upon

consideration of the average number of exons/gene in the species chosen for analysis, upon density of exons that have been called on the genomic sequence, and other empirical rules. Thereafter, one or more among the exons can be chosen for subsequent use in gene expression assay. The putative gene structure is also provided by various of these gene prediction programs.

Where such subsequent gene expression assay uses amplified nucleic acid, considerations such as desired amplicon length, primer synthesis requirements, putative exon length, sequence GC content, existence of possible secondary structure, and the like can be used to identify and select those exons that appear most likely successfully to amplify. Where subsequent gene expression assay relies upon nucleic acid hybridization, whether or not using amplified product, further considerations involving hybridization stringency can be applied to identify that subset of sequences that will most readily permit sequence-specific discrimination at a chosen hybridization and wash stringency. One particular such consideration is avoidance of putative exons that span repetitive sequence; such sequence can hybridize spuriously to nonspecific message, reducing specific signal in the hybridization.

Where the goal is, instead, the identification and confirmation of expression of all, or a plurality, of the exons of a gene, as is desired for detection of alternative splice events, putative exons identified in process 200 can be classified, or binned, bioinformatically into putative genes, as is further described below. Thereafter, all of the exon-

specific exons can be chosen for subsequent confirmation in gene expression assay.

For bioinformatic assay, there are fewer constraints on the sequences that can be tested experimentally, and in this latter case therefore process 300 can output the entirety of the input sequence.

The subset of sequences identified by process 300 as suitable for use in assay is then used in process 400 to create the physical and/or informational substrate for experimental verification of the predictions made in process 200, and thereafter to assay those substrates.

Where the goal is to identify protein coding regions in genomic sequence, the expression of the sequences predicted to encode protein is verified in process 400. The combination of the predictive and experimental methods provides a powerful engine for discovering exons of new genes.

Thus, in another aspect, the present invention provides methods and apparatus for verifying the expression of putative exons identified within genomic sequence. In particular, the invention provides a method of verifying gene expression in which expression of predicted exons is measured and confirmed using a novel type of nucleic acid microarray, the genome-derived single exon nucleic acid microarrays of the present invention.

According to one embodiment of this aspect, predicted exons are amplified from genomic DNA. For generation of gene-specific probes, exons as predicted by a consensus of gene calling, particularly gene prediction, algorithms in process 200, and a- further

identified as suitable by process 300, are amplified. For generation of exon-specific probes useful for detecting splice variants, all predicted exons are typically amplified.

Amplification can be performed using the polymerase chain reaction (PCR). Although PCR is conveniently used, other amplification approaches, such as rolling circle amplification, can also be used.

Amplification schemes can be designed to capture the entirety of each predicted exon in an amplicon with minimal additional (that is, flanking intronic or intergenic) sequence. Because exons predicted from human genomic sequence using the methods of the present invention differ in length, such an approach results in amplicons of varying length.

However, we have found that most exons predicted from human genomic sequence are shorter than 500 bp in length. Although amplicons of at least about 75 base pairs, more preferably at least about 100 base pairs, even more preferably at least about 200 base pairs can be immobilized as probes on nucleic acid microarrays, our early experimental results using the methods of the present invention suggested that longer amplicons, at least about 400 base pairs, more preferably about 500 base pairs, are more effectively immobilized on glass slides or other prepared surfaces.

Although we had suspected that the intronic and intergenic material flanking putative exons in such longer amplicons might cause interference with exon-specific hybridization during microarray experiments, we have found instead, to our surprise, that the ratio of expression of any such probe as between an experimental tissue (or cell type) and a control tissue

is not significantly affected by the presence in the probes of sequence that does not contribute to hybridization to message or cDNA.

Equally surprising, the art had suggested that single exon probes would not provide sufficient signal intensity for high stringency hybridization analyses. Although low stringency hybridization conditions have been designed that permit informative hybridization to highly redundant oligonucleotide-based microarrays, it was believed that the high stringency hybridization conditions typically used for EST-based microarrays would not be usable with single exon probes. We have found, surprisingly, that single-exon probes provide adequate signal at high stringency.

As a result, we have found that we are readily able to use genome-derived amplification products having a single exon flanked by intergenic and/or intronic sequence to confirm the expression of bioinformatically predicted exons.

To the extent that chemical synthesis methods permit oligonucleotides to be generated of sufficient length to encompass an exon, such oligonucleotides can be used as probes in lieu of amplified material. At present, however, amplified products can be generated that exceed the reasonable size limit of chemically synthesized oligonucleotides; amplification thus more readily permits probes to be generated that have single exons flanked by intronic and/or intergenic sequence.

Probes having flanking intergenic and/or intronic sequence permit a wider range of alternative splice events to be detected than do probes that contain only exonic sequence. For example, exon extension would be detectable with such probes as an

increase in signal intensity: we have found a near-linear relationship between signal intensity and length of hybridizing sequence. And when used to assay heteronuclear, *i.e.*, immature mRNA, probes having intronic and/or intergenic flanking sequence permit a wider variety of events to be assessed.

Furthermore, certain advantages derive from application to the microarray of amplicons of defined size.

Therefore, amplification schemes can alternatively, and preferably, be designed to amplify regions of defined size, preferably at least about 300 bp, more preferably at least about 400 bp, most preferably about 500 bp, centered about each predicted exon. Such an approach results in a population of amplicons of limited size diversity, but that typically contain intronic and/or intergenic nucleic acid in addition to, and flanking, the putative exon.

Conversely, somewhat fewer than 10% of exons predicted from human genomic sequence according to the methods of the present invention exceed 500 bp in length. Portions of such longer exons, preferably at least about 300 bp, more preferably at least about 400 bp, most preferably about 500 bp, can be amplified. However, in our early experiments we found that the percentage success at amplifying pieces of such exons is low, and that such putative exons are more effectively amplified when larger fragments, at least about 1000 bp, typically at least about 1500 bp, and even as large as 2000 bp are amplified. Further routine optimization of the PCR reaction would permit 500 bp portions of the longer exons to be amplified.

For amplification, the putative exons selected in process 300 are input into one or more primer design programs, such as PRIMER3 (available online for use at http://www-genome.wi.mit.edu/cgi-bin/primer/), with a goal of amplifying at least about 500 base pairs of genomic sequence centered within or about exons predicted to be no more than about 500 bp (or at least about 1000 - 1500 bp of genomic sequence for exons predicted to exceed 500 bp in length), and the primers synthesized by standard techniques. Primers with the requisite sequences can be purchased commercially or synthesized by standard techniques.

Conveniently, a first predetermined sequence can be added commonly to each exon-specific 5' primer and a second, typically different, predetermined sequence commonly added to each 3' exon-unique primer. This serves to immortalize the amplicon; that is, it serves to permit further amplification of any amplicon using a single set of primers complementary respectively to the common 5' and common 3' sequence elements. The presence of these "universal" priming sequences further facilitates later sequence verification, providing a sequence common to all amplicons at which to prime sequencing reactions. The common 5' and 3' sequences can further serve to add a cloning site should any of the exons warrant further study.

Such predetermined sequence is usefully at least about 10 nt in length, typically at least about 12 nt, more typically about 15 nt in length, and usually does not exceed about 25 nt in length. The "universal" priming sequences used in the examples

presented *infra* were each 16 nt long, and are further described in commonly owned and copending U.S. patent application serial no. 09/608,408, filed June 30, 2000, the disclosure of which is incorporated herein by reference in its entirety.

The genomic DNA to be used as substrate for amplification will come from the eukaryotic species from which the genomic sequence data had originally been obtained, or a closely related species, and can conveniently be prepared by well known techniques from somatic or germline tissue or cultured cells of the organism. See, *e.g.*, Short Protocols in Molecular Biology : A Compendium of Methods from Current Protocols in Molecular Biology, Ausubel et al. (eds.), 4th edition (April 1999), John Wiley & Sons (ISBN: 047132938X) and Maniatis *et al.*, Molecular Cloning : A Laboratory Manual, 2nd edition (December 1989), Cold Spring Harbor Laboratory Press (ISBN: 0879693096), the disclosures of which are incorporated herein by reference in their entireties. Many such prepared genomic DNAs are available commercially, with the human genomic DNAs additionally having certification of donor informed consent.

After partial purification, as by size exclusion spin column or adsorption to glass, with or without confirmation as to amplicon quality as by gel electrophoresis, each amplicon (single exon probe) is disposed in an array upon a support substrate.

Methods for creating microarrays by deposition and fixation of nucleic acids onto support substrates are well known in the art. *Reviewed in* Schena (ed.), DNA Microarrays : A Practical Approach

(Practical Approach Series), Oxford University Press (1999) (ISBN: 0199637768); *Nature Genet.* 21(1)(suppl):1 - 60 (1999); Schena (ed.), Microarray Biochip: Tools and Technology, Eaton Publishing Company/BioTechniques Books Division (2000) (ISBN: 1881299376), the disclosures of which are incorporated herein by reference in their entireties.

Typically, the support substrate will be glass, although other materials, such as amorphous silicon, crystalline silicon, or plastics, can be used. Such plastics include polymethylacrylic, polyethylene, polypropylene, polyacrylate, polymethylmethacrylate, polyvinylchloride, polytetrafluoroethylene, polystyrene, polycarbonate, polyacetal, polysulfone, celluloseacetate, cellulosenitrate, nitrocellulose, or mixtures thereof. Typically, the support will be rectangular, although other shapes, particularly circular disks and even spheres, present certain advantages. Particularly advantageous alternatives to glass slides as support substrates for array of nucleic acids are optical discs, as described in Demers, "Spatially Addressable Combinatorial Chemical Arrays in CD-ROM Format," international patent publication WO 98/12559, incorporated herein by reference in its entirety.

The amplified nucleic acids can be attached covalently to a surface of the support substrate or, more typically, applied to a derivatized surface in a chaotropic agent that facilitates denaturation and adherence by presumed noncovalent interactions, or some combination thereof.

Robotic spotting devices useful for arraying nucleic acids on support substrates can be constructed

using public domain specifications (The MGuide, version 2.0, http://cmgm.stanford.edu/pbrown/mguide/index.html), or can conveniently be purchased from commercial sources (MicroArray GenII Spotter and MicroArray GenIII Spotter, Molecular Dynamics, Inc., Sunnyvale, CA). Spotting can also be effected by printing methods, including those using ink jet technology.

As is well known in the art, microarrays typically also contain immobilized control nucleic acids. For controls useful in providing measurements of background signal for the genome-derived single exon microarrays of the present invention, a plurality of *E. coli* genes can readily be used. As further described in Example 1, 16 or 32 *E. coli* genes suffice to provide a robust measure of nonspecific hybridization in such microarrays.

As is well known in the art, the amplified product disposed in arrays on a support substrate to create a nucleic acid microarray can consist entirely of natural nucleotides linked by phosphodiester bonds, or alternatively can include either nonnative nucleotides, alternative internucleotide linkages, or both, so long as complementary binding can be obtained in the hybridization. If enzymatic amplification is used to produce the immobilized probes, the amplifying enzyme will impose certain further constraints upon the types of nucleic acid analogs that can be generated.

Although particularly described herein as using high density microarrays constructed on planar substrates, the methods of the present invention for confirming the expression of exons predicted from genomic sequence can use any of the known types of

microarrays as herein defined, including microarrays on nonplanar, nonunitary, distributed substrates, such as the nonplanar, bead-based microarrays as are described in Brenner *et al.*, *Proc. Natl. Acad. Sci. USA* 97(4):166501670 (2000); U.S. Patent No. 6,057,107; and U.S. Patent No. 5,736,330, the disclosures of which are incorporated herein by reference in their entireties. In theory, a packed collection of such beads provides in aggregate a higher density of nucleic acid probe than can be achieved with spotting or lithography techniques on a single planar substrate. In addition, gene expression can be confirmed using hybridization to lower density arrays, such as those constructed on membranes, such as nitrocellulose, nylon, and positively-charged derivatized nylon membranes.

Planar microarrays on solid substrates, however, provide certain useful advantages, including compatibility with existing readers. For example, each standard microscope slide can include at least 1000, typically at least 2000, preferably 5000 or more, and up to 19,000 or more nucleic acid probes of discrete sequence.

For purposes of generating gene-specific probes, each putative gene can be represented in the array by a single predicted exon or by a plurality of exons predicted to belong to the same gene. For purposes of generating exon-specific (single exon) probes useful for measuring differential splicing, more than one predicted exon, preferably all predicted exons of a gene, can be tested. And as is well known in the art, each probe of defined sequence, representing a single predicted exon, can be deposited in a plurality

of locations on a single microarray to provide redundancy of signal.

The genome-derived single exon microarrays described above are an important aspect of the present invention, and differ in several fundamental and advantageous ways from microarrays presently used in the gene expression art, including (1) those created by deposition of mRNA-derived nucleic acids, (2) those created by *in situ* synthesis of oligonucleotide probes, and (3) those constructed from yeast genomic DNA.

Most nucleic acid microarrays that are in use for study of eukaryotic gene expression have as immobilized probes nucleic acids that are derived – either directly or indirectly – from expressed message. It is common, for example, for such microarrays to be derived from cDNA/EST libraries, either from those previously described in the literature, such as those from the I.M.A.G.E. consortium, Lennon *et al.*, "The I.M.A.G.E. Consortium: an Integrated Molecular Analysis of Genomes and Their Expression, *Genomics* 33(1):151-2 (1996), or from the *de novo* construction of "problem specific" libraries targeted at a particular biological question, R.S. Thomas *et al.*, *Toxicologist* 54:68-69 (2000). Such microarrays are herein collectively denominated "EST microarrays".

Such EST microarrays by definition can measure expression only of those genes found in EST libraries, which we have recently shown to represent only a fraction of expressed genes. Thus, as further discussed in Example 1, *infra*, and detailed in our commonly owned and copending U.S. provisional patent

application serial nos. 60/180,312, filed February 4, 2000; 60/207,456, filed May 26, 2000; 60/234,687, filed September 21, 2000; 60/236,359, filed September 27, 2000; and U.K. patent application no. 24263.6, filed October 4, 2000, incorporated herein by reference in their entireties, fully 2/3 of genes identified from newly-accessioned human genomic sequence data by the methods of the present invention – and for which expression was subsequently confirmed using the methods and apparatus of the present invention – do not appear in EST or other expression databases, and could not, therefore, have been represented as probes on an EST microarray.

Furthermore, EST and cDNA libraries – and thus microarrays based thereupon – are biased by the tissue or cell type of message origin.

In addition, representation of a message in an EST and/or cDNA library depends upon the successful reverse transcription, optionally but typically with subsequent successful cloning, of the message. This introduces substantial bias into the population of probes available for arraying in EST microarrays. For example, as further demonstrated in the examples, *infra*, the remaining population of genes identified from genomic sequence by the methods of the present invention – that is, the one third of sequences that had previously been accessioned in EST or other expression databases, are biased toward genes with higher expression levels.

In contrast, neither reverse transcription nor cloning is required to produce the probes arrayed on the genome-derived single exon microarrays of the present invention. And although the ultimate

deposition of a probe on the genome-derived single exon microarray of the present invention depends upon a successful amplification from genomic material, *a priori* knowledge of the sequence of the desired amplicon affords greater opportunity to recover any given probe sequence recalcitrant to amplification than is afforded by the requirement for successful reverse transcription and cloning of unknown message in EST approaches. Furthermore, if the sequence cannot be amplified, the sequence can at times be chemically synthesized in its entirety for use in the present invention.

Thus, the genome-derived single exon microarrays of the present invention present a far greater diversity of probes for measuring gene expression, with far less bias, than do EST microarrays presently used in the art.

As a further consequence of their ultimate origin from expressed message, the probes in EST microarrays often contain poly-A (or complementary poly-T) stretches derived from the poly-A tail of mature mRNA. These homopolymeric stretches contribute to cross-hybridization, that is, to a spurious signal occasioned by hybridization to the homopolymeric tail of a labeled cDNA that lacks sequence homology to the gene-specific portion of the probe.

In contrast, the probes arrayed in the genome-derived single exon microarrays of the present invention lack homopolymeric stretches derived from message polyadenylation, and thus can provide more specific signal. Typically, at least about 50% of the probes on the genome-derived single exon microarrays of the present invention lack homopolymeric regions

consisting of A or T, where a homopolymeric region is defined for purposes herein as stretches of 25 or more, typically 30 or more, identical nucleotides. More typically, at least about 60%, even more typically at least about 75%, of probes on the genome-derived single exon microarrays of the present invention lack such homopolymeric stretches.

A further distinction, which also affects the specificity of hybridization, is occasioned by the typical derivation of EST microarray probes from cloned material. Because much of the probe material disposed as probes on EST microarrays is excised or amplified from plasmid, phage, or phagemid vectors, EST microarrays typically include a fair amount of vector sequence, more so when the probes are amplified, rather than excised, from the vector.

In contrast, the vast majority of probes in the genome-derived single exon microarrays of the present invention contain no prokaryotic or bacteriophage vector sequence, having been amplified directly or indirectly from genomic DNA. Typically, therefore, at least about 50%, more typically at least about 60%, 70%, and even 80% or more of individual exon-including probes disposed on a genome-derived single exon microarray of the present invention lack vector sequence, and particularly lack sequences drawn from plasmids and bacteriophage. Preferably, at least about 85%, more preferably at least about 90%, most preferably more than 90% of exon-including probes in the genome-derived single exon microarray of the present invention lack vector sequence. With attention to removal of vector sequences through preprocessing 24, percentages of vector-free exon-including probes

can be as high as 95 - 99%. The substantial absence of vector sequence from the genome-derived single exon microarrays of the present invention results in greater specificity during hybridization, since spurious cross-hybridization to a probe vector sequence is reduced.

As a further consequence of excision or amplification of probes from vectors in construction of EST microarrays, the probes arrayed thereon often contain artificial sequence, derived from vector polylinker multiple cloning sites, at both 5' and 3' ends. The probes disposed upon the genome-derived single exon microarrays need have no such artificial sequence appended thereto.

As mentioned above, however, the exon-specific primers used to amplify putative exons can include artificial sequences, typically 5' to the exon-specific primer sequence, useful for "universal" (that is, independent of exon sequence) priming of subsequent amplification or sequencing reactions. When such "universal" 5' and/or 3' priming sequences are appended to the amplification primers, the probes disposed upon the genome-derived single exon microarray will include artificial sequence similar to that found in EST microarrays. However, the genome-derived single exon microarray of the present invention can be made without such sequences, and if so constructed, presents an even smaller amount of nonspecific sequence that would contribute to nonspecific hybridization.

Yet another consequence of typical use of cloned material as probes in EST microarrays is that such microarrays contain probes that result from cloning artifacts, such as chimeric molecules containing coding region of two separate genes.

Derived from genomic material, typically not thereafter cloned, the probes of the genome-derived single exon microarrays of the present invention lack such cloning artifacts, and thus provide greater specificity of signal in gene expression measurements.

A further consequence of the cloned origin of probes on many EST microarrays is that the individual probes often have disparate sizes, which can cause the optimal hybridization stringency to vary among probes on a single microarray. In contrast, as discussed above, the probes arrayed on the genome-derived single exon microarrays of the present invention can readily be designed to have a narrow distribution in sizes, with the range of probe sizes no greater than about 10% of the average size, typically no greater than about 5% of the average probe size.

Because of their origin from fully- or partially-spliced message, probes disposed upon EST arrays will often include multiple exons. The percentage of such exon-spanning probes in an EST microarray can be calculated, on average, based upon the predicted number of exons/gene for the given species and the average length of the immobilized probes. For human genes, the near-complete sequence of human chromosome 22, Dunham *et al.*, *Nature* 402(6761):489-95 (1999), predicts that human genes average 5.5 exons/gene. Even with probes of 200 - 500 bp, the vast majority of human EST microarray probes include more than one exon.

In contrast, by virtue of their origin from algorithmically identified exons in genomic sequence, the probes in the genome-derived single exon microarrays of the present invention can comprise

individual exons, which provides the ability, as will be further discussed below, to detect and to characterize the expression of splice variants.

Although the presence of multiexon probes will not interfere with the ability to confirm expression of predicted exons in a first level screen, it is preferred that at least about 50%, typically at least about 60%, even more typically at least about 70% of probes disposed on the genome-derived microarray of the present invention consist of, or include, no more than one exon. In preferred embodiments, at least about 75%, more preferably at least about 80%, 85%, 90%, 95%, and even 99% of probes in the genome-derived microarrays of the present invention consist of, or include, no more than one exon.

Although, in the most preferred embodiments, at least about 95%, and even at least about 99% of probes in the genome-derived microarray consist of, or include, no more than one exon, we have found that our early bioinformatic parameters typically produce, at this stage of analysis, about 10% of probes that potentially contain two exons. We expect that some fraction of these probes will prove to encode only a single exon, and that further optimization of our bioinformatic approach will reduce the percentage of probes having more than one potential exon.

Further distinguishing the genome-derived single exon microarrays of the present invention from the EST arrays in the art, the exons that are represented in EST microarrays are often biased toward the 3' or 5' end of their respective genes, since sequencing strategies used for EST identification are so biased. In contrast, no such 3' or 5' bias

necessarily inheres in the selection of exons for disposition on the genome-derived single exon microarrays of the present invention.

Conversely, the probes provided on the genome-derived single exon microarrays of the present invention typically, but need not necessarily, include intronic and/or intergenic sequence that is absent from EST microarrays, which are derived from mature mRNA. As above mentioned, such inclusion, although not mandatory, is advantageous, particularly in use of the probes for detection of alternative splice events. Typically, therefore, at least about 50%, more typically at least about 60%, and even more typically at least about 70% of the exon-including probes on the genome-derived single exon microarrays of the present invention include sequence drawn from noncoding regions. In some embodiments, at least about 80%, more typically at least about 85%, 90%, 91%, 92%, 93%, 94%, 95%, 96%, 97%, 98%, and even 99% or more of exon-including probes on the genome-derived single exon microarrays of the present invention will include sequence drawn from noncoding regions.

The genome-derived single exon microarrays of the present invention are also quite different from *in situ* synthesis microarrays, where probe size is severely constrained by limitations of the photolithographic or other *in situ* synthesis processes.

Typically, probes arrayed on *in situ* synthesis microarrays are limited to a maximum of about 25 bp. As a well known consequence, hybridization to such chips must be performed at low stringency. In order, therefore, to achieve unambiguous sequence-

specific hybridization results, the *in situ* synthesis microarray requires substantial redundancy, with concomitant programmed arraying for each probe of probe analogues with altered (*i.e.*, mismatched) sequence.

In contrast, the longer probe length of the genome-derived single exon microarrays of the present invention allows much higher stringency hybridization and wash. Typically, therefore, exon-including probes on the genome-derived single exon microarrays of the present invention average at least about 100 bp, more typically at least about 200 bp, preferably at least about 250 bp, even more preferably about 300 bp, 400 bp, or in preferred embodiments, at least about 500 bp in length. By obviating the need for substantial probe redundancy, this approach permits a higher density of probes for discrete exons or genes to be arrayed on the microarrays of the present invention than can be achieved for *in situ* synthesis microarrays.

A further distinction is that the probes in *in situ* synthesis microarrays typically are covalently linked to the substrate surface. In contrast, the probes disposed on the genome-derived microarray of the present invention typically are, but need not necessarily be, bound noncovalently to the substrate.

Furthermore, the short probe size on *in situ* microarrays causes large percentage differences in the melting temperature of probes hybridized to their complementary target sequence, and thus causes large percentage differences in the theoretically optimum stringency across the array as a whole.

In contrast, the larger probe size in the microarrays of the present invention create lower

percentage differences in melting temperature across the range of arrayed probes.

A further significant advantage of the microarrays of the present invention over *in situ* synthesized arrays is that the quality of each individual probe can be confirmed before deposition. In contrast, the quality of probes cannot be assessed on a probe-by-probe basis for the *in situ* synthesized microarrays presently being used.

The genome-derived single exon microarrays of the present invention are also distinguished over, and present substantial benefits over, the genome-derived microarrays from lower eukaryotes such as yeast. *See, e.g.*, Lashkari *et al.*, *Proc. Natl. Acad. Sci. USA* 94:13057-13062 (1997).

Only about 220 - 250 of the 6100 or so nuclear genes in *Saccharomyces cerevisiae* — that is, only about 4 to 5% — have standard, spliceosomal, introns, Lopez *et al.*, *Nucl. Acids Res.* 28:85-86 (2000); Spingola *et al.*, *RNA* 5(2):221-34 (1999). Alternative splicing will, therefore, play a far more limited role in such organisms than in higher eukaryotes. A significant aspect of the present invention is the use of exon-specific probes for detecting and characterizing alternative splice events, a role for which whole genome microarrays from lower eukaryotes such as yeast are necessarily ill-suited.

Accordingly, the genome-derived single exon microarrays of the present invention are typically drawn from eukaryotes in which at least about 10%, typically at least about 20%, more typically at least about 50% of protein-encoding genes have introns. In

preferred embodiments, the methods and apparatus of the present invention are used to identify and confirm expression of exons from genomic sequence of eukaryotes in which the average number of introns per gene is at least about one, more typically at least about two, even more typically at least about three or more.

After the physical substrate is prepared, experimental verification of predicted function is performed.

For generating exon-specific probes useful in detection and characterization of alternative splice events, the function sought to be identified in genomic sequence is protein coding and experimental verification of that function is performed by measuring expression of the putative exons. Typically, expression is measured through nucleic acid hybridization experiments, particularly through hybridization to genome-derived single exon microarrays prepared as above-described.

Expression is conveniently measured and reported for each probe in the microarray both as a signal intensity and as a ratio of the expression measured relative to a control, according to techniques well known in the microarray art, *reviewed in* Schena (ed.), DNA Microarrays : A Practical Approach (Practical Approach Series), Oxford University Press (1999) (ISBN: 0199637768); *Nature Genet.* 21(1)(suppl):1 - 60 (1999); Schena (ed.), Microarray Biochip: Tools and Technology, Eaton Publishing Company/BioTechniques Books Division (2000) (ISBN: 1881299376), the disclosures of which are incorporated herein by reference in their entireties. *See also*

Example 2, *infra*. The mRNA source for the reference (control) used to calculate expression ratios can be heterogeneous, as from a pool of multiple tissues and/or cell types or, alternatively, can be drawn from a homogeneous mRNA source, such as a single cultured cell-type.

In Examples 1 and 2, *infra*, we used a pool of 10 tissues/cell types as control. We have since observed that almost every probe that demonstrates expression in the control pool can readily be shown to be expressed in HeLa cells. Since use of a pooled control might mask subtle alternative splice events, we have used HeLa as the source of control message in more recent experiments.

mRNA can be prepared by standard techniques, Short Protocols in Molecular Biology : A Compendium of Methods from Current Protocols in Molecular Biology, Ausubel et al. (eds.), 4th edition (April 1999), John Wiley & Sons (ISBN: 047132938X) and Maniatis *et al.*, Molecular Cloning : A Laboratory Manual, 2nd edition (December 1989), Cold Spring Harbor Laboratory Press (ISBN: 0879693096), the disclosures of which are incorporated herein by reference in their entireties, or purchased commercially. The mRNA is then typically reverse-transcribed in the presence of labeled nucleotides: the index source (that in which expression is desired to be measured) is reverse transcribed in the presence of nucleotides labeled with a first label, typically a fluorophore (equivalently denominated fluorochrome; fluor; fluorescent dye); the reference source is reverse transcribed in the presence of a second label, typically a fluorophore, typically fluorometrically-distinguishable from the first label.

As further described in Example 2, *infra*, Cy3 and Cy5 dyes prove particularly useful in these methods. After partial purification of the index and reference targets, hybridization to the probe array is conducted according to standard techniques, typically under a coverslip or in an automatic slide processing unit.

After wash, microarrays are conveniently scanned using a commercial microarray scanning device, such as a Gen3 or Avalanche Scanner (Molecular Dynamics, Sunnyvale, CA). Data on expression is then passed, with or without interim storage, to process 500, where the results for each probe are related to the original sequence.

Often, hybridization of target material to the genome-derived single exon microarray will identify certain of the probes thereon as of particular interest. Thus, it is often desirable that the user be able readily to obtain sufficient quantities of an individual probe, either for subsequent arrayed deposition upon an additional support substrate, often as part of a microarray having a plurality of probes so identified, or alternatively or additionally as a solitary solid-phase or solution-phase probe, for further use.

Thus, in another aspect, the present invention provides compositions and kits for the ready production of nucleic acids identical in sequence to, or substantially identical in sequence to, probes on the genome-derived single exon microarrays of the present invention.

In one embodiment, the invention provides individual single exon probes in the form of substantially isolated and purified nucleic acid. In

one such embodiment the probe is provided in quantity sufficient to perform a hybridization reaction.

When provided in quantity sufficient to perform a hybridization reaction, the probe can be in any form directly hybridizable to the target that contains the probe's exon (or its complement), such as double stranded DNA, single-stranded DNA complementary to the target, single-stranded RNA complementary to the target, or chimeric DNA/RNA molecules so hybridizable.

The nucleic acid can alternatively or additionally include either nonnative nucleotides, alternative internuclectide linkages, or both, so long as complementary binding can be obtained. For example, probes can include phosphorothioates, methylphosphonates, morpholino analogs, and peptide nucleic acids (PNA), as are described, *inter alia*, in U.S. Patent Nos. 5,142,047; 5,235,033; 5,166,315; 5,217,866; 5,184,444; 5,861,250; international patent applications nos. WO 93/25706; and in *Science* 254:1497 (1991); *J. Am. Chem. Soc.* 114:9677 (1992); *J. Am. Chem. Soc.* 144:1895 (1992); *J. Chem. Soc. Chem. Comm.* 800 (1993); *Proc. Nat. Acad. Sci. USA* 90:1667 (1993); *Intercept Ltd.* 325 (1992); *J. Am. Chem. Soc.* 114:9677 (1992); *Nucleic Acids Res.* 21:197 (1993); *J. Chem. Soc. Chem. Commun.* 518 (1993); *Anti-Cancer Drug Design* 8:53 (1993); *Nucleic Acids Res.* 21:2103 (1993); *Org. Proc. Prep.* 25:457 (1993); CRC Press 363 (1992); *J. Chem. Soc. Chem. Commun.* 9:800 (1993); *J. Am. Chem. Soc.* 115:6477 (1993); *Nature* 365:566 (1993); WO 92/20702; and WO 92/20703, the disclosures of which are incorporated herein by reference.

Usefully, however, such probes are instead provided in a form and quantity suitable for amplification. Typically, such probes are provided in a form and quantity suitable for amplification by PCR or by other well known amplification technique. One such technique additional to PCR is rolling circle amplification, as is described, *inter alia*, in U.S. Patent Nos. 5,854,033 and 5,714,320 and international patent publications WO 97/19193 and WO 00/15779, the disclosures of which are incorporated herein by reference in their entireties. As is well understood, where the probes are to be provided in a form suitable for amplification, the range of nucleic acid analogues and/or internucleotide linkages will be constrained by the requirements and nature of the amplification enzyme.

Where the probe is to be provided in form suitable for amplification, the quantity need not be sufficient for direct hybridization for gene expression analysis, and need be sufficient only to function as an amplification template, typically at least about 1 pg, more typically at least about 10 pg, and usually at least about 100 pg or more.

Each discrete amplifiable probe can also be packaged with amplification primers, either in a single composition that comprises probe template and primers, or in a kit that comprises such primers separately packaged therefrom. As above mentioned, the exon-specific 5' primers used for genomic amplification can have a first common sequence added thereto, and the exon-specific 3' primers used for genomic amplification can have a second, different, common sequence added thereto, thus permitting, in this embodiment, the use

of a single set of 5' and 3' primers to amplify any one of the probes. The probe composition and/or kit can also include buffers, enzyme, *etc.*, required to effect amplification.

In another embodiment, only amplification primers are provided. The primers are sufficient to permit generation of the single exon probe by amplification from genomic DNA, which can be provided by the user.

As mentioned above, when intended for use on a genome-derived single exon microarray of the present invention, the genome-derived single exon probes of the present invention will typically average at least about 75 - 100 bp, more typically at least about 200 bp, preferably at least about 250 bp, even more preferably about 300 bp, 400 bp, or in preferred embodiments, at least about 500 bp in length, including (and typically, but not necessarily centered about) the exon. Furthermore, when intended for use on a genome-derived single exon microarray of the present invention, the genome-derived single exon probes of the present invention will typically not contain a detectable label.

When intended for use in solution phase hybridization, however – that is, for use in a hybridization reaction in which the probe is not first bound to a support substrate (although the target may indeed be so bound) – length constraints that are imposed in microarray-based hybridization approaches will be relaxed, and such probes will typically be labeled.

In such case, the only functional constraint that dictates the minimum size of such probe is that

each such probe must be capable of specifically identifying in a hybridization reaction the exon from which it is drawn. In theory, a probe of as little as 17 nucleotides is capable of uniquely identifying its cognate sequence in the human genome. For hybridization to expressed message – a subset of target sequence that is much reduced in complexity as compared to genomic sequence – even fewer nucleotides are required for specificity.

Therefore, the probes of the present invention can include as few as 20 bp of exon, typically at least about 25 bp of exon, more typically at least about 50 bp or exon, or more. The minimum amount of exon required to be included in the probe of the present invention in order to provide specific signal in either solution phase or microarray-based hybridizations can readily be determined by routine experimentation using standard high stringency conditions.

Such high stringency conditions are described, *inter alia*, in Short Protocols in Molecular Biology : A Compendium of Methods from Current Protocols in Molecular Biology, Ausubel et al. (eds.), 4th edition (April 1999), John Wiley & Sons (ISBN: 047132938X) and Maniatis *et al.*, Molecular Cloning : A Laboratory Manual, 2nd edition (December 1989), Cold Spring Harbor Laboratory Press (ISBN: 0879693096), the disclosures of which are incorporated herein by reference in their entireties. For microarray-based hybridization, standard high stringency conditions can usefully be 50% formamide, 5X SSC, 0.2 µg/µl poly(dA), 0.2 µg/µl human c_ot1 DNA, and 0.5 % SDS, in a humid oven at 42°C overnight, followed by successive washes of the

microarray in 1X SSC, 0.2% SDS at 55°C for 5 minutes, and then 0.1X SSC, 0.2% SDS, at 55°C for 20 minutes. For solution phase hybridization, standard high stringency conditions can usefully be aqueous hybridization at 65°C in 6X SSC. Lower stringency conditions, suitable for cross-hybridization to mRNA encoding structurally- and functionally-related proteins, can usefully be the same as the high stringency conditions but with reduction in temperature for hybridization and washing to room temperature (approximately 25°C).

When intended for use in solution phase hybridization, the maximum size of the single exon probes of the present invention is dictated by the proximity of other expressed exons in genomic DNA: although each single exon probe can include intergenic and/or intronic material contiguous to the exon in the human genome, each probe of the present invention will typically include portions of only one expressed exon.

Thus, each single exon probe will include no more than about 25 kb of contiguous genomic sequence, more typically no more than about 20 kb of contiguous genomic sequence, more usually no more than about 15 kb, even more usually no more than about 10 kb. Usually, probes that are maximally about 5 kb will be used, more typically no more than about 3 kb.

It will be appreciated that single stranded probes must be complementary in sequence to the target; it is well within the skill in the art to determine such complementary sequence and the need therefor. It will further be understood that double stranded probes can be used in both solution-phase hybridization and microarray-based hybridization if suitably denatured.

Thus, it is an aspect of the present invention to provide single-stranded nucleic acid probes that have sequence complementary to those described herein above and below, and double-stranded probes one strand of which has sequence complementary to the probes described herein.

As mentioned above, the probes can, but need not, contain intergenic and/or intronic material that flanks the exon, on one or both sides, in the same linear relationship to the exon that the intergenic and/or intronic material bears to the exon in genomic DNA. The probes typically do not, however, contain nucleic acid derived from more than one expressed exon.

And when intended for use in solution hybridization, the probes of the present invention can usefully have detectable labels. Nucleic acid labels are well known in the art, and include, *inter alia*, radioactive labels, such as ^{3}H, ^{32}P, ^{33}P, ^{35}S, ^{125}I, ^{131}I; fluorescent labels, such as Cy3, Cy5, Cy5.5, Cy7, SYBR® Green and other labels described in Haugland, *Handbook of Fluorescent Probes and Research Chemicals*, 7th ed., Molecular Probes Inc., Eugene, OR (2000), or fluorescence resonance energy transfer tandem conjugates thereof; labels suitable for chemiluminescent and/or enhanced chemiluminescent detection; labels suitable for ESR and NMR detection; quantum dots; and labels that include one member of a specific binding pair, such as biotin, digoxigenin, or the like.

The probes, either in quantity sufficient for hybridization or sufficient for amplification, can be provided in individual vials or containers, and can be

provided dry (e.g., lyophilized), or solvated. If solvated, the solution can usefully include buffers and salts as desired for hybridization and/or amplification. Furthermore, if desired to be spotted on a microarray, the probes can usefully be provided in a solution of chaotropic agent to facilitate adherence to the microarray support substrate.

Alternatively, such probes can usefully be packaged as a plurality of such individual genome-derived single exon probes.

In one embodiment of this aspect, a small quantity of each probe is disposed, typically without attachment to substrate, in a spatially-addressable ordered set, typically one per well of a microtiter dish. Although a 96 well microtiter plate can be used, greater efficiency is obtained using higher density arrays, such as are provided by microtiter plates having 384, 864, 1536, 3456, 6144, or 9600 wells. And although microtiter plates having physical depressions (wells) are conveniently used, any device that permits addressable withdrawal of reagent from fluidly-noncommunicating areas can be used.

Each of the probes of the ordered set can be provided in any of the forms that are described above with respect to the probes as individually packaged.

As above mentioned, the exon-specific 5' primers used for genomic amplification can have a first common sequence added thereto, and the exon-specific 3' primers used for genomic amplification can have a second, different, common sequence added thereto, thus permitting, in certain embodiments, the use of a single set of 5' and 3' primers to amplify any one of the probes from the amplifiable ordered set.

In another aspect of the present invention, a genome-derived single-exon microarray is packaged together with an addressable set of individual probes, the set of individual probes including at least a subset of the probes on the microarray. In alternative embodiments, the ordered set of amplifiable probes is packaged separately from the genome-derived single exon microarray.

In some embodiments, the microarray and/or ordered probe set are further packaged with recorded media that provide probe identification and addressing information, and that can additionally contain annotation information, such as gene expression data. Such recorded media can be packaged with the microarray, with the ordered probe set, or with both.

If the microarray is constructed on a substrate that incorporates recordable media, such as is described in international patent application no. WO 98/12559, entitled "Spatially addressable combinatorial chemical arrays in CD-ROM format," incorporated herein by reference in its entirety, then separate packaging of the genome-derived single exon microarray and the bioinformatic information is not required.

Although the use of high density genome-derived microarrays on solid planar substrates is presently a preferred approach for the physical confirmation and characterization of the expression of sequences predicted to encode protein, other types of microarrays, as well as lower density macro arrays, can also be used.

Experimental verification in process 400 of the function predicted from genomic sequence in process

200 can be bioinformatic, rather than, or additional to, physical verification.

Where the function desired to be identified is protein coding, the predicted exons can be compared bioinformatically to sequences known or suspected of being expressed.

Thus, the sequences output from process 300 (or process 200), can be used to query expression databases, such as EST databases, SNP ("single nucleotide polymorphism") databases, known cDNA and mRNA sequences, SAGE ("serial analysis of gene expression") databases, and more generalized sequence databases that allow query for expressed sequences. Such query can be done by any sequence query algorithm, such as BLAST ("basic local alignment search tool"). The results of such query — including information on identical sequences and information on nonidentical sequences that have diffuse or focal regions of sequence homology to the query sequence — can then be passed directly to process 500, or used to inform analyses subsequently undertaken in process 200, process 300, or process 400.

Experimental data, whether obtained by physical or bioinformatic assay in process 400, is passed to process 500 where it is usefully related to the sequence data itself, a process termed "annotation". Such annotation can be done using any technique that usefully relates the functional information to the sequence, as, for example, by incorporating the functional data into the record itself, by linking records in a hierarchical or relational database, by linking to external databases,

or by a combination thereof. Such database techniques are well within the skill in the art.

The annotated sequence data can be stored locally, uploaded to genomic sequence database 100, and/or displayed 800.

The above-described methods and apparatus are extremely efficient means for identifying exons in raw genomic sequence data and confirming their expression. As described in further detail in commonly owned and copending U.S. provisional application nos. 60/207,456, filed May 26, 2000; 60/234,687, filed September 21, 2000; 60/236,359, filed September 27, 2000; and U.K. patent application no. 24263.6, filed October 4, 2000, the disclosures of which are incorporated herein by reference in their entirety, we have used this process to identify more than 15,000 exons in human genomic sequence whose expression we have confirmed in at least one human tissue or cell type. Fully two-thirds of the exons belong to genes that were not then represented in existing public expression (EST, cDNA) databases.

Accordingly, the above-described methods and apparatus are an extremely efficient way to generate exon-specific probes for high throughput identification and characterization of alternative splice events, as will hereinafter be described. Other methods, however, can be used to design exon-specific probes for use in the methods of the present invention for detecting and characterizing alternative splice events.

For example, the spliced sequence alignment method of Gelfand *et al.*, *Proc. Natl. Acad. Sci. USA* 93:9061-9066 (1996) can be used to identify exon-intron structure in genomic sequence data. From such results, putative exons can be amplified and their expression

confirmed by hybridization as above-described. Methods particularly designed to predict alternative splicing, such as those set forth in Hanke *et al.*, *Trends Genet.* 15(10):389-390 (1999), Mironov *et al.*, *Genome Res.* 9:1288-93 (1999), and Croft *et al.*, *Nature Genet.* 24:340-341 (2000), will also prove useful in design of exon-specific probes, as will the more classical *ad hoc* anecdotal approaches that compare cDNA structures for individual genes to their respective genomic sequences.

Increasingly, use of these various methods will lead to annotation of general sequence databases with such exon-specific information and will, in parallel, lead to increased representation of genes and exons in specialized databases devoted to alternative splicing, such as are described, *inter alia*, in Gelfand *et al.*, *Nucl. Acids Res.* 27(1):301-302 (1999); Dralyuk *et al.*, *Nucl. Acids Res.* 28(1):296-297 (2000); Croft *et al.* Increasingly, therefore, it will be possible to design exon-specific probes for use in the methods of the present invention for detecting and characterizing splice variants by query of such databases. As noted above, however, to the extent that such methods rely upon mRNA, cDNA or EST sequence, they will suffer from various constraints that do not attend the methods of the present invention for designing exon-specific probes.

Display of Annotated Genomic Sequence

The methods and apparatus of the present invention rapidly produce functional information from genomic sequence. Coupled with the escalating pace at which sequence now accumulates, the rapid pace of

sequence annotation produces a need for methods of displaying the information in meaningful ways. It is, therefore, another aspect of the present invention to provide means for displaying annotated sequence, and in particular for displaying sequence annotated with exon-specific expression information, according to the methods and apparatus of the present invention. Further, such display can be used as a preferred graphical user interface for electronic search, query, and analysis of such annotated sequence.

FIG. 3 schematizes visual display 80 presenting a single genomic sequence annotated according to the present invention. Because of its nominal resemblance to artistic works of Piet Mondrian, visual display 80 is alternatively described herein as a "Mondrian".

Each of the visual elements of display 80 is aligned with respect to the genomic sequence being annotated (the "annotated sequence"). Given the number of nucleotides typically represented in an annotated sequence, representation of individual nucleotides would rarely be readable in hard copy output of display 80. Typically, therefore, the annotated sequence is schematized as rectangle 89, extending from the left border of display 80 to its right border. By convention herein, the left border of rectangle 89 represents the first nucleotide of the sequence and the right border of rectangle 89 represents the last nucleotide of the sequence.

As further discussed below, however, the Mondrian visual display of annotated sequence can serve as a convenient graphical user interface for computerized representation, analysis, and query of

information stored electronically. For such use, the individual nucleotides can conveniently be linked to the X axis coordinate of rectangle 89. This permits the annotated sequence at any point within rectangle 89 readily to be viewed, either automatically — for example, by time-delayed appearance of a small overlaid window ("tool tip") upon movement of a cursor or other pointer over rectangle 89 — or through user intervention, as by clicking a mouse or other pointing device at a point in rectangle 89.

Visual display 80 is generated after user specification of the genomic sequence to be displayed. Such specification can consist of or include an accession number for a single clone (e.g., a single BAC accessioned into GenBank), wherein the starting and stopping nucleotides are thus absolutely identified, or alternatively can consist of or include an anchor or fulcrum point about which a chosen range of sequence is anchored, thus providing relative endpoints for the sequence to be displayed. For example, the user can anchor such a range about a given chromosomal map location, gene name, or even a sequence returned by query for similarity or identity to an input query sequence. When visual display 80 is used as a graphical user interface to computerized data, additional control over the first and last displayed nucleotide will typically be dynamically selectable, as by use of standard zooming and/or selection tools.

Field 81 of visual display 80 is used to present the output from process 200, that is, to present the bioinformatic prediction of those sequences having the desired function within the genomic sequence. Functional sequences are typically indicated

by at least one rectangle 83 (83a, 83b, 83c), the left and right borders of which respectively indicate, by their X-axis coordinates, the starting and ending nucleotides of the region predicted to have function.

Where a single bioinformatic method or approach identifies a plurality of regions having the desired function, a plurality of rectangles 83 is disposed horizontally in field 81. Where multiple methods and/or approaches are used to identify function, each such method and/or approach can be represented by its own series of horizontally disposed rectangles 83, each such horizontally disposed series of rectangles offset vertically from those representing the results of the other methods and approaches.

Thus, rectangles 83a in FIG. 3 represent the functional predictions of a first method of a first approach for predicting function, rectangles 83b represent the functional predictions of a second method and/or second approach for predicting that function, and rectangles 83c represent the predictions of a third method and/or approach.

Where the function desired to be identified is protein coding, field 81 is used to present the bioinformatic prediction of sequences encoding protein. For example, rectangles 83a can represent the results from GRAIL or GRAIL II, rectangles 83b can represent the results from GENEFINDER, and rectangles 83c can represent the results from DICTION.

Optionally, and preferably, rectangles 83 collectively representing predictions of a single method and/or approach are identically colored and/or textured, and are distinguishable from the color and/or texture used for a different method and/or approach.

Alternatively, or in addition, the color, hue, density, or texture of rectangles 83 can be used further to report a measure of the bioinformatic reliability of the prediction. For example, many gene prediction programs will report a measure of the reliability of prediction. Thus, increasing degrees of such reliability can be indicated, *e.g.*, by increasing density of shading. Where display 80 is used as a graphical user interface, such measures of reliability, and indeed all other results output by the program, can additionally or alternatively be made accessible through linkage from individual rectangles 83, as by time-delayed window ("tool tip" window), or by pointer (*e.g.*, mouse)-activated link.

As above described, increased predictive reliability can be achieved by requiring consensus among methods and/or approaches to determining function. Thus, field 81 can include a horizontal series of rectangles 83 that indicate one or more degrees of consensus in predictions of function, including the combined length of the separately predicted exons that overlap in frame.

Although FIG. 3 shows three series of horizontally disposed rectangles in field 81, display 80 can include as few as one such series of rectangles and as many as can discriminably be displayed, depending upon the number of methods and/or approaches used to predict a given function. For example, addition of a fourth gene prediction program, such as GENSCAN (http://genes.mit.edu/GENSCANinfo.html), to the three gene prediction programs used in our first experiments (GRAIL, GENEFINDER, DICTION) would be accommodated by a

fourth series of rectangles disposed horizontally in field 81, but offset vertically from rectangles 81a, 81b, and 81c.

Furthermore, field 81 can be used to show predictions of a plurality of different functions. However, the increased visual complexity occasioned by such display makes more useful the ability of the user to select a single function for display. When display 80 is used as a graphical user interface for computer query and analysis, such function can usefully be indicated and user-selectable, as by a series of graphical buttons or tabs (not shown in FIG. 3).

Rectangle 89 is shown in FIG. 3 as including interposed rectangle 84. Rectangle 84 represents the portion of annotated sequence for which predicted functional information has been assayed physically, with the starting and ending nucleotides of the assayed material indicated by the X axis coordinates of the left and right borders of rectangle 84. Rectangle 85, with optional inclusive circles 86 (86a, 86b, and 86c) displays the results of such physical assay.

Although a single rectangle 84 is shown in FIG. 3, physical assay is not limited to just one region of annotated genomic sequence. It is expected that an increasing percentage of regions predicted to have function by process 200 will be assayed physically, and that display 80 will accordingly, for any given genomic sequence, have an increasing number of rectangles 84 and 85, representing an increased density of sequence annotation. For example, for purposes of generating exon-specific probes for alternative splice detection, it is preferred that a plurality of exons, preferably all of the exons, that

commonly belong to a single gene will be assayed experimentally for expression; accordingly, display 80 will have, for the genomic sequence encompassing such exons, a series of rectangles 84 and 85 for each of the assayed exons.

Where the function desired to be identified is protein coding, rectangle 84 identifies the sequence of the probe used to measure expression. In embodiments of the present invention where expression is measured using genome-derived single exon microarrays, rectangle 84 identifies the sequence included within the probe immobilized on the support surface of the microarray. As noted *supra*, such probe will often include a small amount of additional, synthetic, material incorporated during amplification and designed to permit reamplification of the probe, which sequence is typically not shown in display 80.

Rectangle 87 is used to present the results of bioinformatic assay of the genomic sequence. For example, where the function desired to be identified is protein coding, process 400 can include bioinformatic query of expression databases with the sequences predicted in process 200 to encode exons. And as above discussed, because bioinformatic assay presents fewer constraints than does physical assay, often the entire output of process 200 can be used for such assay, without further subsetting thereof by process 300. Therefore, rectangle 87 typically need not have separate indicators therein of regions submitted for bioinformatic assay; that is, rectangle 87 typically need not have regions therein analogous to rectangles 84 within rectangle 89.

Rectangle 87 as shown in FIG. 3 includes smaller rectangles 880 and 88. Rectangles 880 indicate regions that returned a positive result in the bioinformatic assay, with rectangles 88 representing regions that did not return such positive results. Where the function desired to be predicted and displayed is protein coding, rectangles 880 indicate regions of the predicted exons that identify sequence with significant similarity in expression databases, such as EST, SNP, SAGE databases, with rectangles 88 indicating genes novel over those identified in existing expression data bases.

Rectangles 880 can further indicate, through color, shading, texture, or the like, additional information obtained from bioinformatic assay.

For example, where the function assayed and displayed is protein coding, the degree of shading of rectangles 880 can be used to represent the degree of sequence similarity found upon query of expression databases. The number of levels of discrimination can be as few as two (identity, and similarity, where similarity has a user-selectable lower threshold). Alternatively, as many different levels of discrimination can be indicated as can visually be discriminated.

Where display 80 is used as a graphical user interface, rectangles 880 can additionally provide links directly to the sequences identified by the query of expression databases, and/or statistical summaries thereof. As with each of the precedingly-discussed uses of display 80 as a graphical user interface, it should be understood that the information accessed via display 80 need not be resident on the computer

presenting such display, which often will be serving as a client, with the linked information resident on one or more remotely located servers.

Rectangle 85 displays the results of physical assay of the sequence delimited by its left and right borders.

Rectangle 85 can consist of a single rectangle, thus indicating a single assay, or alternatively, and increasingly typically, will consist of a series of rectangles (85a, 85b, 85c) indicating separate physical assays of the same sequence.

Where the function assayed is gene expression, and where gene expression is assayed as herein described using simultaneous two-color fluorescent detection of hybridization to genome-derived single exon microarrays, individual rectangles 85 can be colored to indicate the degree of expression relative to control. Conveniently, shades of green can be used to depict expression in the sample over control values, and shades of red used to depict expression less than control, corresponding to the spectra of the Cy3 and Cy5 dyes conventionally used for respective labeling thereof. Additional functional information can be provided in the form of circles 86 (86a, 86b, 86c), where the diameter of the circle can be used to indicate a parameter different from that set forth in rectangle 85. For example, where the annotated function is protein coding, rectangle 85 can report expression relative to control and circle 86 can be used to report signal intensity. As discussed *infra*, such relative expression (expression ratio) and absolute expression (signal intensity) can be expressed using normalized values.

Where display 80 is used as a graphical user interface, rectangle 85 can be used as a link to further information about the assay. For example, where the assay is one for gene expression, each rectangle 85 can be used to link to information about the source of the hybridized mRNA, the identity of the control, raw or processed data from the microarray scan, or the like.

For purposes of illustration only, FIG. 4 shows an embodiment of display 80 showing typical color conventions when hypothetical genomic sequence is annotated with exon-specific expression data. As would of course readily be understood, the color choice is arbitrary, and alternative colors can be used.

In this typical presentation, BAC sequence ("Chip seq.") 89 is presented in red, with the physically assayed region thereof (corresponding to rectangle 84 in FIG. 3) shown in white. Algorithmic gene predictions are shown in field 81, with predictions by GRAIL shown in green, predictions by GENEFINDER shown in blue, and predictions by DICTION shown in pink. Within rectangle 87, regions of sequence that, when used to query expression databases, return identical or similar sequences ("EST hit") are shown as white rectangles (corresponding to rectangles 880 in FIG. 3), gray indicates low homology, and black indicates unknowns (where black and gray would correspond to rectangles 88 in FIG. 3).

Although FIGS. 3 and 4 show a single stretch of sequence, uninterrupted from left to right, longer sequences are usefully represented by vertical stacking of such individual Mondrians, as shown in FIGS. 5 and 6.

Associating Exons That Belong to a Common Gene

To detect and characterize splice variants, it is preferred that exon-specific probes be available to interrogate expression of a plurality of exons, preferably all of the potentially-expressed exons, of a given gene. It is, therefore, another aspect of the present invention to provide methods for associating the exons, as predicted from genomic sequence, that contribute to a single gene.

For the third of genes identified by the methods of the present invention that are identically present in expression databases, the prior-accessioned data can be used to associate the predicted exons and can, further, also be used to order the exons. As noted above, however, the fragmentary nature of much of such EST-based data will often preclude conclusive association of all of the exons of a given gene. Furthermore, exons that appear uniquely or predominantly in splice variants with restricted expression — whether restricted temporally, as during development, or restricted spatially, as in rare cell types, or restricted environmentally, as in cells expressing variants under rare environmental conditions — will be under-represented in such expression databases and cannot, therefore, readily be associated by such means.

The 3' and 5' end bias of EST libraries (and thus of the data in EST databases), however, will prove useful in such circumstances, since often the 5'-most extreme and 3'-most extreme exons will be present in such fragmentary data. Exons identified in genomic

sequence lying between such extreme exons will presumptively, albeit not conclusively, belong to a single gene.

For the third of genes identified by the methods of the present invention that are not identically present in expression databases but that have adequate sequence similarity to genes known to be expressed, reference to the structure of the known genes can usefully suggest which exons be associated. Such comparisons of course suffer the limitations of any EST-based comparison.

For the third of genes that are neither identically present nor similar to genes known to be expressed, as well as for the other 2/3 of genes above-discussed, proximity of exons within genomic sequence will be useful in associating exons that belong to a common gene. Although proximity in genomic sequence is useful, its utility for associating exons is limited by the difficulty of discerning the 5' and 3' boundaries of any given gene, particularly since the size of the genomic locus that encodes any given eukaryotic gene can vary quite dramatically.

Additional criteria, such as the presence in genomic sequence of recognizable promoter elements or sequence elements that are associated with transcriptional or translational starts will prove useful in anchoring the 5' ends of genes. Analogously, identification of sequence elements that are associated with 3' transcript processing will be useful in anchoring the 3' ends of genes. However, Mironov *et al., supra*, classifying predicted transcript alternatives by mRNA region, report that 80% of the predicted alternatively spliced genes had an

alternative in the 5'-untranslated region (5'-UT), 19% had alternatives in the 3'-untranslated region (3'-UT), and 20% had alternatives in the coding region, suggesting that exact predictions as to 5'-most and 3'-most exons from genomic sequence will be difficult.

Three of the four gene prediction (exon-finding) programs that we use — GENEFINDER, GENSCAN, and GRAIL — also predict gene structure, including promoters, initial, internal, and terminal exons, and frame information. It is possible, therefore, to use a consensus among these three programs, or between any two of the three, to predict gene structure-- that is, to predict that exons predicted in genomic sequence commonly belong to a single gene.

Using our visual display tool, the Mondrian, we have found that consensus in the pattern of expression of individual exons is a powerful means for identifying exons that commonly belong to a single gene. It is, therefore, another aspect of the present invention to provide methods, including methods based upon visual display, for associating exons that commonly belong to a single gene using, as the criterion for association, consensus in their patterns of expression in a plurality of tissues and/or cell types.

As further discussed in Example 3, FIG. 5 presents a Mondrian of BAC AC008172 (bases 25,000 to 130,000 shown), containing the carbamyl phosphate synthetase gene (AF154830.1), the sequence and structure of which has previously been reported. Purple background within the region shown as field 81 in FIG. 3 indicates all 37 known exons for this gene.

As can be seen, GRAIL II successfully identified 27 of the known exons (73%), GENEFINDER successfully identified 37 of the known exons (100%), while DICTION identified 7 of the known exons (19%).

Seven of the predicted exons were selected for physical assay, of which 5 successfully amplified by PCR and were sequenced. These five exons were all found to be from the same gene, the carbamyl phosphate synthetase gene (AF154830.1).

The five exons were arrayed and gene expression measured across 10 tissues. As is readily seen by visual inspection of the resulting Mondrian (FIG. 5), the five single-exon probes report identical expression ratio patterns: each exon is expressed above control (i.e., in green) in the tissues represented by the fourth, seventh, and eighth rectangles (corresponding to rectangles 85 in FIG. 3) and is expressed at or below control in the remaining tissues.

Of course, an exon that is removed or truncated by alternatively splicing in one of the assayed tissues would produce a variant expression pattern. For purposes of associating exons as belonging commonly to a single gene, however, a consensus among assayed tissues would still identify the exon as presumptively belonging to the same gene.

The methods of this aspect of the invention can, and typically will, be automated. For example, WO 99/58720, incorporated herein by reference in its entirety, describes algorithms for ordering the relatedness of a plurality of multidimensional expression data sets. The methods set forth therein can readily be adapted to ordering the relatedness of data sets, wherein each data set comprises expression

ratios of an individual exon across a plurality of tissues and cell types, permitting exons with related, but not necessarily identical, patterns of expression to be classified as belonging to a common gene.

Use of Genome-Derived Single Exon Probes in High Throughput Interrogation of Exon-Specific Expression

The genome-derived single exon probes and genome-derived single exon microarrays of the present invention permit the high throughput interrogation of myriad tissues and cell types for exon-specific expression as further described and exemplified, *e.g.*, in Examples 1 and 2, *infra*. Where a plurality of exons from a single gene are so assayed, the exon-specific expression data from plural exons of a single gene permit a variety of alternative splice events to be detected and characterized, as further described below and exemplified in Example 4, *infra*.

The plural exon-specific probes need not be disposed adjacent to one another on a microarray, nor need they even be disposed on the same microarray, to provide the necessary data. There are advantages, however, to disposing plural single exon probes specific for different exons of a single gene on a single microarray: such approach permits the contemporaneous measurement of expression from each of the plural exons of a gene in a single hybridization experiment, and reduces error.

Thus, in another aspect, the invention provides genome-derived single exon microarrays for which at least about 10%, more typically at least about 20%, even more typically at least about 30%, yet even

more typically at least about 40% or even at least about 50% of the genes having a first exon present among the probes disposed thereon have at least a second exon present among the single exon probes of the microarray. For maximal efficiency, typically at least about 60%, more typically at least 70%, even more typically at least about 75%, and for maximal efficiency, at least about 80% or more of the genes having a first exon present among the probes disposed thereon have at least a second exon present among the single exon probes of the microarray.

In particularly useful embodiments, on average at least about 25% of the exons are present for each gene represented on the microarray. For increased efficiency, at least about 30% of the exons are present for each gene represented, even more efficiently at least about 40% of the exons are present, even more so at least about 50% of the exons are present, and for maximal efficiency, at least about 60% of the exons are present, on average, for each gene represented on the microarray, with higher efficiencies achieved with minimal average exon presence of 70%, 80%, or even 90%, 95% and, at the limit, 100% of the exons present for each gene represented on the microarray.

The genome-derived single exon microarrays that are particularly adapted to detection of alternative splicing, as just described, otherwise advantageously have the properties described above for the genome-derived single exon microarrays used to confirm exon expression, which description is incorporated here by reference. Thus, by way of example, typically at least about 50% of the probes lack homopolymeric regions consisting of A or T, where

a homopolymeric region is defined for purposes herein as stretches of 25 or more, typically 30 or more, identical nucleotides. More typically, at least about 60%, even more typically at least about 75%, of probes on the genome-derived single exon microarrays of the present invention lack such homopolymeric stretches.

Further by way of example, genome-derived single exon arrays particularly adapted for detection of splice variants can, like those used to confirm expression of predicted exons, have lower probe density and be constructed, *e.g.*, on membranes, such as nitrocellulose, nylon, and positively-charged derivatized nylon membranes. Further, such arrays can be nonplanar, bead-based microarrays such as are described in Brenner *et al.*, *Proc. Natl. Acad. Sci. USA* 97(4):166501670 (2000); U.S. Patent No. 6,057,107; and U.S. Patent No. 5,736,330, the disclosures of which are hereby incorporated herein by reference in their entireties.

Also by way of illustration, on the microarrays particularly adapted to splice detection, advantageously at least about 50%, more typically at least about 60%, and even more typically at least about 70% of the exon-including probes on the genome-derived single exon microarray include sequence drawn from noncoding regions. In some embodiments, at least about 80%, more typically at least about 90% of exon-including probes on the genome-derived single exon microarrays of the present invention will include sequence drawn from noncoding regions.

The inclusion of noncoding region is not required: as discussed above, exon-specific expression can be measured using probes having only exonic

sequence. However, the inclusion of noncoding region expands the variety of alternative splice events that can be detected. For example, Mironov *et al., supra,* report that a high percentage of alternative splice events create alternatives in the 5' and 3' UT. By including, within the probes having the 5' and 3' most-extreme exons of a gene, portions respectively of the 5'-UT and 3'-UT, alternatives in the 5'-UT and 3'-UT can be detected by changes in hybridization to these probes.

As noted above, which discussion is incorporated here by reference, the present invention provides compositions and kits for the ready production of nucleic acids identical in sequence to, or substantially identical in sequence to, probes on the genome-derived single exon microarrays of the present invention. And as further noted *supra,* which discussion is also incorporated here by reference, the invention further provides probes packaged as a plurality of such individual genome-derived single exon probes.

Accordingly, the invention provides individual single exon probes and plural subsets of probes that correspond to those disposed upon the genome-derived single exon microarrays particularly adapted to splice detection.

In particular, the invention provides sets of plural single exon probes for which at least about 10%, more typically at least about 20%, even more typically at least about 30%, yet even more typically at least about 40% or even at least about 50% of the genes having a first exon present among the probes in the set have at least a second exon present among the probes in

the set. For maximal efficiency, typically at least about 60%, more typically at least 70%, even more typically at least about 75%, and for even greater efficiency, at least about 80% or more of the genes having a first exon present among the probes have at least a second exon also present among the single exon probes of the set.

In particularly useful embodiments, on average at least about 25% of the exons are present for each gene represented by at least one exon in the probe set. For increased efficiency, at least about 30% of the exons are present for each represented gene, even more efficiently at least about 40% of the exons are present, even more so at least about 50% of the exons are present, and for maximal efficiency, at least about 60% of the exons are present, on average, for each gene represented in the probe set, with higher efficiencies achieved with minimal average exon presence of 70%, 80%, or even 90%, 95% and, at the limit, 100% of the exons present for each gene represented in the probe set.

For detection of splice variants, exon-specific expression will be assayed in a variety of different tissues and cell types. Among such tissues and cell types are normal differentiated tissues and cell types of adults, cell types and tissues that normally appear transiently during development, and abnormal and/or diseased tissues and cell types. In Example 4, *infra*, placenta, fetal liver, skeletal muscle, prostate, liver, lung, kidney, HeLa, heart, brain, bone marrow, and adrenal were assayed for expression.

Sources of tissue and cell types are readily available and include the American Type Culture Collection (Manassas, VA), and, at the Coriell Institute for Medical Research, the NIGMS Human Genetic Cell Repository, the NIA Aging Cell Repository, the Autism Research Resource, the ADA Cell Repository Maturity Onset Diabetes Collection, and the HBDI Cell Repository Juvenile Diabetes Collection. Others are well known in the art.

Furthermore, cells and tissues from genotypically disparate, yet phenotypically normal, individuals can be assessed, permitting genotypic changes (SNPs) to be related to differentially expressed splice variants.

Although probe density can be quite high on the genome-derived single exon microarrays of the present invention, it will be infrequent that all exons of a eukaryotic genome can be disposed on a single array. Typically, subsets of exons will be so chosen.

For example, exons can be chosen for inclusion on the splice-detecting microarray ("splice chip") based upon the sequence relatedness of their respective genes (*e.g.*, exons from various G-protein coupled receptors). Exons can be chosen for inclusion commonly on a single splice chip based upon expression commonly in a known developmental time point and/or tissue or cell type, based upon known or suspected relatedness (homology) to genes known to be spliced in other organisms (*e.g.*, drosophila, *C. elegans*, zebra fish, mouse, rat, chimp, ape), based on known drug targets, and based upon known biological pathways.

Example 4, *infra*, demonstrates that simultaneous two-color hybridization experiments using

genome-derived microarrays having plural exons of a gene, each disposed within a discrete single exon probe, generates expression data that readily permit alternative splice events to be detected and characterized in a high throughput screen.

Use of Alternative Genome-Derived Probes in High Throughput Interrogation of Exon-Specific Expression

As described above and exemplified in Example 4, *infra*, the genome-derived single exon probes of the present invention can be used to generate exon-specific expression data suitable for detecting alternative splice events.

However, the single-exon probes as above-described are illustrative, not exhaustive, of the types of probes that can be designed from genomic sequence data and that can be used to interrogate exon-specific expression by high throughput microarray analysis. Thus, in other aspects, the invention provides methods and apparatus for detecting alternative splicing using other types of genome-derived probes.

In a first such aspect, the invention provides methods and apparatus for detecting alternative splicing using overlapping series of genome-spanning probes.

Following identification, *e.g.* by the above-described methods, of a plurality of exons that belong to a common gene, a genomic region encompassing all of the exons is specified. The region can, and typically will, include sequence that lies 5' to the 5' most-extreme exon and sequence that lies 3' to the 3'-most

extreme exon. The greater the extent of upstream and downstream sequence included, the greater the number and types of alternative splice events that can be identified. Typically, the flanking sequence will not be so extensive, however, as to encompass exons from adjacent genes. A series of overlapping probes is then designed that span the entire identified genomic locus.

As would be understood, decreasing probe length and increasing overlap will each independently increase the number of probes required to span the genomic region. For survey purposes, longer probes with lesser overlap will typically be used; for in-depth analysis, shorter probes having greater overlap will typically be used.

Several types of genes will typically warrant analysis using genome-spanning probes. For example, genes for which one or more alternative splice events have previously been detected using the single exon probes of the present invention, as above-described, will usefully be subject to more in-depth analysis using the genome-spanning probes here described. Other genes that usefully will be assessed using genome-spanning probes will be those that have been shown to be associated with disease, and particularly those for which alternative splice forms have been shown to be associated with disease.

For example, Klamt *et al.*, *Hum. Molec. Genet.* 7(4):709-714 (1998) report that Frasier syndrome is caused by defective alternative splicing of *WT1*. For example, Qi *et al.*, *Hum. Molec. Genet.* 7(3):465-469 (1998) report that constitutive skipping of alternatively spliced exon 10 in the *ATP7A* gene

produces occipital horn syndrome. *See also* Vidal-Puig *et al.*, *J. Clin. Invest.* 99:2416-2422 (1997) and Brett *et al.*, *FEBS Letters* 474:83-86 (2000).

Genes that have single nucleotide polymorphisms (SNPs) present in a substantial percentage of one or more populations are also preferred candidates for such in-depth examination, particularly where the SNP alters a potential splice donor or acceptor sequence. Genes that are homologues of genes known to be spliced in other species, or that are homologues of genes associated with disease in other species, are also usefully probed for alternative splice events using genome-spanning probes.

Typically, for disposition on a microarray, the probes will be at least about 75 base pairs, more preferably at least about 100 base pairs, even more preferably at least about 200 base pairs. Because our early experimental results suggested that longer amplicons are more effectively immobilized, at least about 400 base pairs, more preferably about 500 base pairs, will be used.

As with the single exon probes of the invention, such overlapping genome-spanning probes will usefully have a first primer sequence appended to the 5' end and a second, different, primer sequence appended to the 3' end, which primer sequences will be identical as among the probes.

To detect expression, the probes are disposed on a nucleic acid microarray, as defined herein, and used to measure expression by hybridization. Usefully, such hybridization can be done using simultaneous two-color fluorescent hybridization techniques as are well

known in the art and exemplified, *e.g.*, in Examples 1 and 2, *infra*.

We have observed that the intensity of hybridization signal is approximately linearly proportional to the length of the hybridizing portion of the probe on the microarray. Thus, probes overlapping an exon will typically produce a graded intensity series, with the leading and lagging edge probes reporting lesser signal intensity than the central probes having greater exon-specific sequence.

In a second such aspect, the invention provides methods and apparatus for detecting alternative splicing using genome-derived probes comprising untranslated sequences.

As noted above, the methods of the present invention can be used to identify regions of genomic sequence that are predicted to have any biological function specified by the user, not just protein coding. Accordingly, comparative sequence analysis can be used in process 200 (see FIGS. 1 and 2) to identify regions of genomic sequence that have low interspecies variability – that is, are strongly conserved – but that nonetheless may not meet bioinformatic criteria (*e.g.*, as above-described) for encoding protein. Process 10 of the present invention can thereafter be used, as above-described, to generate genome-derived probes that contain the identified low variability sequence. Disposing such probes on a genome-derived microarray and hybridizing to message-derived nucleic acids (as described herein with respect to genome-derived single exon probes) permit a wide variety of alternative splice events to be detected in the 5'-UT

and 3'-UT that might not be detectable using the single exon probes and microarrays of the present invention.

Visual Display of Exon-Specific Expression Measurements

In another aspect, the present invention presents methods and apparatus for displaying exon-specific measurements in a format that facilitates detection of alternative splicing events.

In one embodiment, the visual display has the attributes of a Mondrian, as above-described, but the X-axis excludes sequence not included in the amplicons. Alternatively, the X axis excludes all sequence not present in exons, thus rendering the X axis presentation equivalent to the presumed mRNA sequence. Further, in such visual display the gene predictions that, in a Mondrian, are set forth in rectangles 83 can usefully be omitted. Examples are shown in FIGS. 12 and 13, which are further described in Example 4.

The following examples are offered by way of illustration and not by way of limitation.

EXAMPLE 1
Preparation of Single Exon Microarrays from Exons Predicted in Human Genomic Sequence

Bioinformatics Results

All human BAC sequences in fewer than 10 pieces that had been accessioned in a five month period immediately preceding this study were downloaded from

GenBank. This corresponds to ≈2200 clones, totaling ≈350 MB of sequence, or approximately 10% of the human genome.

After masking repetitive elements using the program CROSS_MATCH, the sequence was analyzed for open reading frames using three separate gene finding programs. The three programs predict genes using independent algorithmic methods developed on independent training sets: GRAIL uses a neural network, GENEFINDER uses a hidden Markoff model, and DICTION, a program proprietary to Genetics Institute, operates according to a different heuristic. The results of all three programs were used to create a prediction matrix across the segment of genomic DNA.

The three gene finding programs yielded a range of results. GRAIL identified the greatest percentage of genomic sequence as putative coding region, 2% of the data analyzed. GENEFINDER was second, calling 1%, and DICTION yielded the least putative coding region, with 0.8% of genomic sequence called as coding region.

The consensus data were as follows. GRAIL and GENEFINDER agreed on 0.7% of genomic sequence, GRAIL and DICTION agreed on 0.5% of genomic sequence, and the three programs together agreed on 0.25% of the data analyzed. That is, 0.25% of the genomic sequence was identified by all three of the programs as containing putative coding region.

Exons predicted by any two of the three programs ("consensus exons") were assorted into "gene bins" using two criteria: (1) any 7 consecutive exons within a 25 kb window were placed together in a bin as likely contributing to a single gene, and (2) all exons

within a 25 kb window were placed together in a bin as likely contributing to a single gene if fewer than 7 exons were found within the 25 kb window.

PCR

The largest exon from each gene bin that did not span repetitive sequence was then chosen for amplification, as were all consensus exons longer than 500 bp. This method approximated one exon per gene; however, a number of genes were found to be represented by multiple elements.

Previously, we had determined that DNA fragments fewer than 250 bp in length do not bind well to the amino-modified glass surface of the slides used as support substrate for construction of microarrays; therefore, amplicons were designed in the present experiments to approximate 500 bp in length.

Accordingly, after selecting the largest exon per gene bin, a 500 bp fragment of sequence centered on the exon was passed to the primer picking software, PRIMER3 (available online for use at http://www-genome.wi.mit.edu/cgi-bin/primer/). A first additional sequence was commonly added to each exon-unique 5' primer, and a second, different, additional sequence was commonly added to each exon-unique 3' primer, to permit subsequent reamplification of the amplicon using a single set of "universal" 5' and 3' primers, thus immortalizing the amplicon. The addition of universal priming sequences also facilitates sequence verification, and can be used to add a cloning site should some exons be found to warrant further study.

The exons were then PCR amplified from genomic DNA, verified on agarose gels, and sequenced using the universal primers to validate the identity of the amplicon to be spotted in the microarray.

Primers were supplied by Operon Technologies (Alameda, CA). PCR amplification was performed by standard techniques using human genomic DNA (Clontech, Palo Alto, CA) as template. Each PCR product was verified by SYBR® green (Molecular Probes, Inc., Eugene, OR) staining of agarose gels, with subsequent imaging by Fluorimager (Molecular Dynamics, Inc., Sunnyvale, CA). PCR amplification was classified as successful if a single band appeared.

The success rate for amplifying exons of interest directly from genomic DNA using PCR was approximately 75%. FIG. 7 graphs the distribution of predicted exon length and distribution of amplified PCR products, with exon length shown by dashed line and PCR product length shown by solid line. Although the range of exon sizes is readily seen to extend to beyond 900 bp, the mean predicted exon size was only 229 bp, with a median size of 150 bp (n=9498). With an average amplicon size of 475 ± 25 bp, approximately 50% of the average PCR amplification product contained predicted coding region, with the remaining 50% of the amplicon containing either intron, intergenic sequence, or both.

Using a strategy predicated on amplifying about 500 bp, it was found that long exons had a higher PCR failure rate. To address this, the bioinformatics process was adjusted to amplify 1000, 1500 or 2000 bp fragments from exons larger than 500 bp. This improved the rate of successful amplification of exons exceeding

500 bp, constituting about 9.2% of the exons predicted by the gene finding algorithms.

Approximately 75% of the probes disposed on the array (90% of those that successfully PCR amplified) were sequence-verified by sequencing in both the forward and reverse direction using MegaBACE sequencer (Molecular Dynamics, Inc., Sunnyvale, CA), universal primers, and standard protocols.

Some genomic clones (BACs) yielded very poor PCR and sequencing results. The reasons for this are unclear, but may be related to the quality of early draft sequence or the inclusion of vector and host contamination in some submitted sequence data.

Although the intronic and intergenic material flanking coding regions could theoretically interfere with hybridization during microarray experiments, subsequent empirical results demonstrated that differential expression ratios were not significantly affected by the presence of noncoding sequence. The variation in exon size was similarly found not to affect differential expression ratios significantly; however, variation in exon size was observed to affect the absolute signal intensity (data not shown).

The 350 MB of genomic DNA was, by the above-described process, reduced to 9750 discrete probes, which were spotted in duplicate onto glass slides using commercially available instrumentation (MicroArray GenII Spotter and/or MicroArray GenIII Spotter, Molecular Dynamics, Inc., Sunnyvale, CA). Each slide additionally included either 16 or 32 *E. coli* genes, the average hybridization signal of which was used as a measure of background biological noise.

Each of the probe sequences was BLASTed against the human EST data set, the NR data set, and SwissProt GenBank (May 7, 1999 release 2.0.9).

One third of the probe sequences (as amplified) produced an exact match (BLAST Expect ("E") values less than 1 e^{-100}) to either an EST (20% of sequences) or a known mRNA (13% of sequences). A further 22% of the probe sequences showed some homology to a known EST or mRNA (BLAST E values from 1 e^{-5} to 1 e^{-99}). The remaining 45% of the probe sequences showed no significant sequence homology to any expressed, or potentially expressed, sequences present in public databases.

All of the probe sequences (as amplified) were then analyzed for protein similarities with the SwissProt database using BLASTX, Gish *et al.*, *Nature Genet.* 3:266 (1993). The predicted functional breakdowns of the 2/3 of probes identical or homologous to known sequences are presented in Table 1.

Table 1

Function of Predicted Exons As Deduced From Comparative Sequence Analysis			
Total	V6 chip	V7 chip	Function Predicted from Comparative Sequence Analysis
211	96	115	Receptor
120	43	77	Zinc Finger
30	11	19	Homeobox
25	9	16	Transcription Factor
17	11	7	Transcription
118	57	61	Structural
95	39	56	Kinase
36	18	18	Phosphatase
83	31	52	Ribosomal
45	19	26	Transport
21	7	14	Growth Factor
17	12	5	Cytochrome
50	33	17	Channel

As can be seen, the two most common types of genes were transcription factors and receptors, making up 2.2% and 1.8% of the arrayed elements, respectively.

EXAMPLE 2

Gene Expression Measurements From Genome-Derived Single Exon Microarrays

The two genome-derived single exon microarrays prepared according to Example 1 were hybridized in a series of simultaneous two-color fluorescence experiments to (1) Cy3-labeled cDNA synthesized from message drawn individually from each of brain, heart, liver, fetal liver, placenta, lung, bone marrow, HeLa, BT 474, or HBL 100 cells, and (2) Cy5-labeled cDNA prepared from message pooled from all ten tissues and cell types, as a control in each of the measurements. Hybridization and scanning were carried

out using standard protocols and Molecular Dynamics equipment.

Briefly, mRNA samples were bought from commercial sources (Clontech, Palo Alto, CA and Amersham Pharmacia Biotech (APB)). Cy3-dCTP and Cy5-dCTP (both from APB) were incorporated during separate reverse transcriptions of 1 µg of polyA+ mRNA performed using 1 µg oligo(dT)12-18 primer and 2 µg random 9mer primers as follows. After heating to 70°C, the RNA:primer mixture was snap cooled on ice. After snap cooling on ice, added to the RNA to the stated final concentration was: 1X Superscript II buffer, 0.01 M DTT, 100µM dATP, 100 µM dGTP, 100 µM dTTP, 50 µM dCTP, 50 µM Cy3-dCTP or Cy5-dCTP 50 µM, and 200 U Superscript II enzyme. The reaction was incubated for 2 hours at 42°C. After 2 hours, the first strand cDNA was isolated by adding 1 U Ribonuclease H, and incubating for 30 minutes at 37°C. The reaction was then purified using a Qiagen PCR cleanup column, increasing the number of ethanol washes to 5. Probe was eluted using 10 mM Tris pH 8.5.

Using a spectrophotometer, probes were measured for dye incorporation. Volumes of both Cy3 and Cy5 cDNA corresponding to 50 pmoles of each dye were then dried in a Speedvac, resuspended in 30 µl hybridization solution containing 50% formamide, 5X SSC, 0.2 µg/µl poly(dA), 0.2 µg/µl human c_0t1 DNA, and 0.5 % SDS.

Hybridizations were carried out under a coverslip, with the array placed in a humid oven at 42°C overnight. Before scanning, slides were washed in 1X SSC, 0.2% SDS at 55°C for 5 minutes, followed by 0.1X SSC, 0.2% SDS, at 55°C for 20 minutes. Slides

were briefly dipped in water and dried thoroughly under a gentle stream of nitrogen.

Slides were scanned using a Molecular Dynamics Gen3 scanner, as described. Schena (ed.), Microarray Biochip: Tools and Technology, Eaton Publishing Company/BioTechniques Books Division (2000) (ISBN: 1881299376).

Although the use of pooled cDNA as a reference permitted the survey of a large number of tissues, it attenuates the measurement of relative gene expression, since every highly expressed gene in the tissue/cell type-specific fluorescence channel will be present to a level of at least 10% in the control channel. Because of this fact, both signal and expression ratios (the latter hereinafter, "expression" or "relative expression") for each probe were normalized using the average ratio or average signal, respectively, as measured across the whole slide.

Data were accepted for further analysis only when signal was at least three times greater than biological noise, the latter defined by the average signal produced by the *E. coli* control genes.

The relative expression signal for these probes was then plotted as a function of tissue or cell type, and is presented in FIG. 8.

FIG. 8 shows the distribution of expression across a panel of ten tissues. The graph shows the number of sequence-verified products that were either not expressed ("0"), expressed in one or more but not all tested tissues ("1" - "9"), and expressed in all tissues tested ("10").

Of 9999 arrayed elements on the two microarrays (including positive and negative controls

and "failed" products), 2353 (51%) were expressed in at least one tissue or cell type. Of the gene elements showing significant signal – where expression was scored as "significant" if the normalized Cy3 signal was greater than 1, representing signal 5-fold over biological noise (0.2) – 39% (991) were expressed in all 10 tissues. The next most common class (15%) consisted of gene elements expressed in only a single tissue.

The genes expressed in a single tissue were further analyzed, and the results of the analyses are compiled in FIG. 9.

FIG. 9A is a matrix presenting the expression of all verified sequences that showed signal intensity greater than 3 in at least one tissue. Each clone is represented by a column in the matrix. Each of the 10 tissues assayed is represented by a separate row in the matrix, and relative expression (expression ratio) of a clone in that tissue is indicated at the respective node by intensity of green shading, with the intensity legend shown in panel B. The top row of the matrix ("EST Hit") contains "bioinformatic" rather than "physical" expression data – that is, presents the results returned by query of EST, NR and SwissProt databases using the probe sequence. The legend for "bioinformatic expression" (*i.e.*, degree of homology returned) is presented in panel C. Briefly, white is known, black is novel, with gray depicting nonidentical with significant homology (white: E values < 1e-100; gray: E values from 1e-05 to 1e-99; black: E values > 1e-05).

As FIG. 9 readily shows, heart and brain were demonstrated to have the greatest numbers of genes that

were shown to be uniquely expressed in the respective tissue. In brain, 200 uniquely expressed genes were identified; in heart, 150. The remaining tissues gave the following figures for uniquely expressed genes: liver, 100; lung, 70; fetal liver, 150; bone marrow, 75; placenta, 100; HeLa, 50; HBL, 100; and BT474, 50.

It was further observed that there were many more "novel" genes among those that were up-regulated in only one tissue, as compared with those that were down-regulated in only one tissue. In fact, it was found that exons whose expression was measurable in only a single of the tested tissues were represented in sequencing databases at a rate of only 11%, whereas 36% of the exons whose expression was measurable in 9 of the tissues were present in public databases. As for those exons expressed in all ten tissues, fully 46% were present in existing expressed sequence databases. These results are not unexpected, since genes expressed in a greater number of tissues have a higher likelihood of being, and thus of having been, discovered by EST approaches.

Comparison of Signal from Known and Unknown Genes

The normalized signal of the genes found to have high homology to genes present in the GenBank human EST database were compared to the normalized signal of those genes not found in the GenBank human EST database. The data are shown in FIG. 10.

FIG. 10 shows in red the normalized Cy3 signal intensity for all sequence-verified products with a BLAST Expect ("E") value of greater than 1e-30

(designated "unknown") upon query of existing EST, NR and SwissProt databases, and shows in blue the normalized Cy3 signal intensity for all sequence-verified products with a BLAST Expect value of less than 1e-30 ("known"). Note that biological background noise has an averaged normalized Cy3 signal intensity of 0.2.

As expected, the most highly expressed of the exons were "known" genes. This is not surprising, since very high signal intensity correlates with very commonly-expressed genes, which have a higher likelihood of being found by EST sequence.

However, a significant point is that a large number of even the high expressers were "unknown". Since the genomic approach used to identify genes and to confirm their expression does not bias toward either the 3' or 5' end of a gene, many of these high expression genes will not have been detected in an end-sequenced cDNA library.

The significant point is that presence of the gene in an EST database is not a prerequisite for incorporation into a genome-derived microarray, and further, that arraying such "unknown" exons can help to assign function to as-yet undiscovered genes.

Verification of Gene Expression

To ascertain the validity of the approach described above to identify genes from raw genomic sequence, expression of two of the probes was assayed using reverse transcriptase polymerase chain reaction (RT PCR) and northern blot analysis.

Two microarray probes were selected on the basis of exon size, prior sequencing success, and

tissue-specific gene expression patterns as measured by the microarray experiments. The primers originally used to amplify the two respective exons from genomic DNA were used in RT PCR against a panel of tissue-specific cDNAs (Rapid-Scan gene expression panel 24 human cDNAs) (OriGene Technologies, Inc., Rockville, MD).

Sequence AL079300_1 was shown by microarray hybridization to be present in cardiac tissue, and sequence AL031734_1 was shown by microarray experiment to be present in placental tissue (data not shown). RT-PCR on these two sequences confirmed the tissue-specific gene expression as measured by microarrays, as ascertained by the presence of a correctly sized PCR product from the respective tissue type cDNAs.

Clearly, all microarray results cannot, and indeed should not, be confirmed by independent assay methods, or the high throughput, highly parallel advantages of microarray hybridization assays will be lost. However, in addition to the two RT-PCR results presented above, the observation that 1/3 of the arrayed genes exist in expression databases provides powerful confirmation of the power of our methodology – which combines bioinformatic prediction with expression confirmation using genome-derived single exon microarrays – to identify novel genes from raw genomic data.

To verify that the approach further provides correct characterization of the expression patterns of the identified genes, a detailed analysis was performed of the microarrayed sequences that showed high signal in brain.

For this latter analysis, sequences that showed high (normalized) signal in brain, but which showed very low (normalized) signal (less than 0.5, determined to be biological noise) in all other tissues, were further studied. There were 82 sequences that fit these criteria, approximately 2% of the arrayed elements. The 10 sequences showing the highest signal in brain in microarray hybridizations are detailed in Table 2, along with assigned function, if known or reasonably predicted.

Table 2

Function of the Most Highly Expressed Genes Expressed Only in Brain				
Microarray Reference Name	Normalized Signal	Expression Ratio	Homology to EST present in GenBank	Gene Function as described by GenBank
AP000217-1	5.2	+ 7.7	High	S-100 protein, b-chain, Ca^{2+} binding protein expressed in central nervous system
AP000047-1	2.3		High	Unknown Function
AC006548-9	1.7		High	Similar to mouse membrane glyco-protein M6, expressed in central nervous system
AC007245-5	1.5		High	Similar to amphiphysin, a synaptic vesicle-associated protein. Ref 21

Function of the Most Highly Expressed Genes Expressed Only in Brain				
Microarray Sequence Name	Normalized Signal	Expression Ratio	Homology to EST present in GenBank	Gene Function as described by GenBank
L44140-4	1.2	+ 2.0	High	Endothelial actin-binding protein found in nonmuscle filamin
AC004689-9	1.2	+ 3.5	High	Protein Phosphatase PP2A, neuronal/ downregulates activated protein kinases
AL031657-1	1.2	+ 3.0	High	Unknown function/ Contains the anhyrin motif, a common protein sequence motif
AC009266-2	1.1	+ 3.7	Low	Low homology to the Synaptotagmin I protein in rat/present at low levels throughout rat brain
AP000086-1	1.0	+ 2.7	Low	Unknown, very poor homology to collagen
AC004689-3	1.0		High	Protein Phosphatase PP2A, neuronal/ downregulates activated protein kinases

Of the ten sequences studied by these latter confirmatory approaches, eight were previously known. Of these eight, six had previously been reported to be important in the central nervous system or brain. The exon giving the highest signal (AP00217-1) was found to be the gene encoding an S100B Ca^{2+} binding protein, reported in the literature to be highly and uniquely expressed in the central nervous system. Heizmann, *Neurochem. Res.* 9:1097 (1997).

A number of the brain-specific probe sequences (including AC006548-9, AC009266-2) did not have homology to any known human cDNAs in GenBank but did show homology to rat and mouse cDNAs. Sequences AC004689-9 and AC004689-3 were both found to be phosphatases present in neurons (Millward *et al.*, *Trends Biochem. Sci.* 24(5):186-191 (1999)). Two microarray sequences, AP000047-1 and AP000086-1 have unknown function, with AP000086-1 being absent from GenBank. Functionality can now be narrowed down to a role in the central nervous system for both of these genes, showing the power of designing microarrays in this fashion.

Next, the function of the chip sequences with the highest (normalized) signal intensity in brain, regardless of expression in other tissues, was assessed. In this latter analysis, we found expression of many more common genes, since the sequences were not limited to those expressed only in brain. For example, looking at the 20 highest signal intensity spots in brain, 4 were similar to tubulin (AC00807905; AF146191-2; AC007664-4; AF14191-2), 2 were similar to actin (AL035701-2; AL034402-1), and 6 were found to be homologous to glyceraldehyde-3-phosphate dehydrogenase

(GAPDH) (AL035604-1; Z86090-1; AC006064-L, AC006064-K; AC035604-3; AC006064-L). These genes are often used as controls or housekeeping genes in microarray experiments of all types.

Other interesting genes highly expressed in brain were a ferritin heavy chain protein, which is reported in the literature to be found in brain and liver (Joshi *et al.*, *J. Neurol. Sci.* 134(Suppl):52-56 (1995)), a result confirmed with the array. Other highly expressed chip sequences included a translation elongation factor 1α (AC007564-4), a DEAD-box homolog (AL023804-4), and a Y-chromosome RNA-binding motif (Chai *et al.*, *Genomics* 49(2):283-89 (1998))(AC007320-3). A low homology analog (AP00123-1/2) to a gene, DSCR1, thought to be involved in trisomy 21 (Down's syndrome), showed high expression in both brain and heart, in agreement with the literature (Fuentes *et al.*, *Mol. Genet.* 4(10):1935-44 (1995)).

As a further validation of the approach, we selected the BAC AC006064 to be included on the array. This BAC was known to contain the GAPDH gene, and thus could be used as a control for the exon selection process. The gene finding and exon selection algorithms resulted in choosing 25 exons from BAC AC006064 for spotting onto the array, of which four were drawn from the GAPDH gene. Table 3 shows the comparison of the average expression ratio for the 4 exons from BAC006064 compared with the average expression ratio for 5 different dilutions of a commercially available GAPDH cDNA (Clontech).

Table 3

Comparison of Expression Ratio, for each tissue, of GAPDH		
	AC006064 (n = 4)	Control (n = 5)
Bone Marrow	-1.81 ± 0.11	-1.85 ± 0.08
Brain	-1.41 ± 0.11	-1.17 ± 0.05
BT474	1.85 ± 0.09	1.66 ± 0.12
Fetal Liver	-1.62 ± 0.07	-1.41 ± 0.05
HBL100	1.32 ± 0.05	2.64 ± 0.12
Heart	1.16 ± 0.09	1.56 ± 0.10
HeLa	1.11 ±0.06	1.30 ± 0.15
Liver	-1.62 ± 0.22	-2.07 ±
Lung	-4.95 ± 0.93	-3.75 ± 0.21
Placenta	-3.56 ± 0.25	-3.52 ± 0.43

Each tissue shows excellent agreement between the experimentally chosen exons and the control, again demonstrating the validity of the present exon mining approach. In addition, the data also show the variability of expression of GAPDH within tissues, calling into question its classification as a housekeeping gene and utility as a housekeeping control in microarray experiments.

EXAMPLE 3
Representation of Sequence and Expression Data as a "Mondrian"

For each genomic clone processed for microarray as above-described, a plethora of information was accumulated, including full clone sequence, probe sequence within the clone, results of

each of the three gene finding programs, EST information associated with the probe sequences, and microarray signal and expression for multiple tissues, challenging our ability to display the information.

Accordingly, we devised a new tool for visual display of the sequence with its attendant annotation which, in deference to its visual similarity to the paintings of Piet Mondrian, is hereinafter termed a "Mondrian". FIGS. 3 and 4 present the key to the information presented on a Mondrian.

FIG. 5 presents a Mondrian of BAC AC008172 (bases 25,000 to 130,000 shown), containing the carbamyl phosphate synthetase gene (AF154830.1). Purple background within the region shown as field 81 in FIG. 3 indicates all 37 known exons for this gene. As can be seen, GRAIL II successfully identified 27 of the known exons (73%), GENEFINDER successfully identified 37 of the known exons (100%), while DICTION identified 7 of the known exons (19%).

Seven of the predicted exons were selected for physical assay, of which 5 successfully amplified by PCR and were sequenced. These five exons were all found to be from the same gene, the carbamyl phosphate synthetase gene (AF154830.1).

The five exons were arrayed, and gene expression measured across 10 tissues. As is readily seen in the Mondrian, the five chip sequences on the array show identical expression patterns, elegantly demonstrating the reproducibility of the system.

FIG. 6 is a Mondrian of BAC AL049839. We selected 12 exons from this BAC, of which 10 successfully sequenced, which were found to form between 5 and 6 genes. Interestingly, 4 of the genes

on this BAC are protease inhibitors. Again, these data elegantly show that exons selected from the same gene show the same expression patterns, depicted below the red line. From this figure, it is clear that our ability to find known genes is very good. A novel gene is also found from 86.6 kb to 88.6 kb, upon which all the exon finding programs agree. We are confident we have two exons from a single gene since they show the same expression patterns and the exons are proximal to each other. Backgrounds in the following colors indicate a known gene (top to bottom): red = kallistatin protease inhibitor (P29622); purple = plasma serine protease inhibitor (P05154); turquoise = α1 anti-chymotrypsin (P01011); mauve = 40S ribosomal protein (P08865). Note that chip sequence 8 and 12 did not sequence verify.

EXAMPLE 4
Use of Genome-Derived Single Exon Microarrays to Detect Alternative Splice Events

We have detected alternative splice events in both known and unknown genes using genome-derived single exon microarrays. One instance of each is presented here.

About 10 megabases of human genomic sequence data was bioinformatically processed essentially as set forth in Examples 1 and 2, with the following exceptions.

Binning of exons, with selection of one representative exon per putative gene, was not performed. Instead, all putative exons called by any of the gene prediction programs and that did not overlap were input into the primer picking software.

Where overlapping exons were called, one was selected for input into the primer picking software.

All successfully amplified amplicons (≈ 3,000 probes) were deposited on a glass slide as described in Examples 1 and 2, *supra*, and used in simultaneous two-color hybridization experiments, essentially as described in Examples 1 and 2. Tissues probed were placenta, fetal liver, skeletal muscle, prostate, liver, lung, kidney, HeLa heart, brain, bone marrow, and adrenal; control in the experiments was provided by HeLa. Exemplary hybridization of the microarray to kidney is shown in two views in FIG. 11.

FIG. 12 is a splice viewer showing a subset of data from these experiments.

For this view, data from 13 of the >3000 probes were selected. The probes include discrete exons of the human nuclear aconitase mitochondrial protein, a known gene. We determined that the particular probes include discrete exons from the same gene by observing an identical sequence match of the respective exons within the amplicons with the known sequence of the complete mRNA.

Nucleotide positions of the mature mRNA are shown in 500 bp increments at the top. Exons are delimited by vertical bars. Those exons present in single exon probes on the microarray are numbered from 1 - 13. Because the microarray did not include all known exons of the gene, the numbers do not necessarily coincide with exon number in the mature mRNA; the numbers do, however, coincide with the SEQ ID NO of the amplicons set forth in the accompanying Sequence Listing, incorporated herein by reference in its

entirety. Missing exons are shown by a red bar in the row above the placenta expression row.

As further described in the legend of FIG. 12, expression ratios relative to control are shown in varying shades of red and green, with insignificant difference in expression from control shown in white. Shown in gray are those data points for which data were unavailable.

As can be seen, the consensus expression pattern for exons of this gene is increased expression (green) in both skeletal muscle and heart muscle tissue, as compared to expression in HeLa (the control).

Deviating from the consensus pattern, exons 1 and 4 are not upregulated in skeletal muscle and heart, suggesting that these exons are either truncated or spliced out in these two tissues.

Also deviating from the consensus pattern, exon 11 is not upregulated in skeletal muscle, relative to control, but is present in heart (green), indicating that this exon is spliced out in skeletal muscle and remains present in heart.

These experiments thus reveal a minimum of three putative forms of the aconitase gene. Form one is the full length gene, as described previously and as present in GenBank. Form two is expressed in heart and lacks (or has truncated versions of) exons included in amplicons 1 and 4. Form three, expressed in skeletal muscle, lacks (or has truncated versions of) exons included in amplicons 1, 4, and 11. Noise in the experiment does not allow conclusions to be drawn regarding the exons included within amplicons 5 and 8.

A second example, drawn from the same experimental data and shown in the splice viewer presented in FIG. 13, demonstrates alternative splicing of a previously unknown gene.

FIG. 13 depicts an unknown gene and putative splice events occurring within the transcript. Exons are numbered consecutively to correspond to the SEQ ID NOs. of the amplicons that contain the exons.

We predict that amplicons 14, 15, 16, and 17 all include discrete exons from the same novel gene based both on the proximity of the amplicon sequences within the human genome (all within 100 kbases) and on the consensus expression patterns observed in the microarray experiments.

As judged by signal intensity and expression [illegible] data, the exons within amplicons 14, 15, 16 and 17 are expressed in kidney and HeLa.

As can be seen, there are two forms of this gene. Form one consists of exons contained in amplicons 14, 15, 16 and 17. Form 2 consists exons covered by amplicons 14, 15, and 17. Form 1 is expressed in HeLa cells and form 2 is expressed in kidney.

Furthermore, the exon within amplicon 14 is solely expressed in skeletal muscle, indicating that this amplicon is either included in a second gene that is expressed in skeletal muscle or, in the alternative, is included in a splice variant expressed only (among the tested tissues) in skeletal muscle.

All patents, patent publications, and other published references mentioned herein are hereby incorporated by reference in their entireties as if each had been individually and specifically incorporated by reference herein. While preferred illustrative embodiments of the present invention are described, one skilled in the art will appreciate that the present invention can be practiced by other than the described embodiments, which are presented for purposes of illustration only and not by way of limitation. The present invention is limited only by the claims that follow.

What is claimed is:

1. A method of detecting alternatively spliced mRNA variants, comprising:

detecting variations, as among a plurality of tissues or cell types, in the expression pattern of a plurality of a gene's exons, said expression pattern measured for each of said tissues or cell types by concurrent hybridization of a plurality of exon-specific probes to transcript-derived nucleic acids from said tissue or cell type,

wherein each of said plurality of exon-specific probes includes a fragment of no more than one exon of a eukaryotic genome, said fragment selectively hybridizable at high stringency to an expressed gene,

wherein said probes collectively include specifically hybridizable fragments of a plurality of exons of at least one gene,

wherein said genome has, on average, at least one intron per gene, and

wherein said plurality of probes averages at least 100 bp in length.

2. The method of claim 1, wherein said plurality of exon-specific probes are disposed, prior to said hybridization, on a nucleic acid microarray.

3. The method of claim 2, wherein at least 50% of probes in said plurality of exon-specific probes include, contiguous to a first end of said probe's included exon fragment, a first intronic and/or intergenic sequence that is identically contiguous to said fragment in said genome.

4. The method of claim 2, wherein at least 50% of probes in said plurality of exon-specific probes include, contiguous to a first end of said probe's included exon fragment, a first intronic and/or intergenic sequence that is identically contiguous to said exon in said genome, and further comprise, contiguous to a second end of said exon fragment, a second intronic and/or intergenic sequence that is identically contiguous to said exon fragment in said genome.

5. The method of claim 2, wherein at least 50% of probes in said plurality of exon-specific probes include at least a first priming sequence not found in contiguity with the rest of said probe sequence in said genome.

6. The method of claim 2, wherein said eukaryotic genome is a human genome.

7. The method of claim 2, wherein the average size of probes in said plurality of exon-specific probes is at least about 200 bp.

8. The method of claim 2, wherein the average size of probes in said plurality of exon-specific probes is at least about 300 bp.

9. The method of claim 2, wherein the average size of probes in said plurality of exon-specific probes is about 500 bp.

10. The method of claim 2, wherein at least 50% of probes in said plurality of exon-specific probes lack prokaryotic and bacteriophage vector sequence.

11. The method of claim 2, wherein at least 50% of probes in said plurality of exon-specific probes lack homopolymeric stretches of A or T.

12. The method of claim 2, wherein at least 50% of the genes having a first exon fragment present among the probes in said plurality of probes have at least a second exon fragment present among the probes.

13. The method of claim 2, wherein at least 50% of the exons are detectable for each gene represented in said plurality.

14. The method of claim 2, wherein each of said plurality of probes is derived, directly or indirectly, by amplification of genomic DNA.

15. A method of producing exon-specific probes useful for detecting alternative splicing, comprising:

predicting an exon from genomic sequence data;

amplifying from genomic DNA a portion thereof that contains said predicted exon and no other predicted exon; and then

confirming by hybridization of said amplified portion that the predicted exon is expressed in a transcribed message.

16. A nucleic acid microarray for detecting alternatively spliced mRNA variants, comprising:

a plurality of nucleic acid probes addressably disposed upon a substrate,

wherein at least 50% of said nucleic acid probes include a fragment of no more than one exon of a eukaryotic genome, said fragment selectively hybridizable at high stringency to an expressed gene,

wherein said plurality of nucleic acid probes averages at least 100 bp in length,

wherein said eukaryotic genome averages at least one intron per gene, and

wherein said probes collectively include fragments of a plurality of exons of at least one gene.

17. The nucleic acid microarray of claim 16, wherein said fragment includes at least 20 contiguous nucleotides of said exons.

18. The nucleic acid microarray of claim 16, wherein said fragment includes at least 25 contiguous nucleotides of said exons.

19. The nucleic acid microarray of claim 16, wherein said fragment includes the entirety of said exon.

20. The nucleic acid microarray of claim 16, wherein at least 50% of said exon-including nucleic acid probes further comprise, contiguous to a first end of said fragment, a first intronic and/or intergenic sequence that is identically contiguous to said fragment in said eukaryotic genome.

21. The nucleic acid microarray of claim 16, wherein at least 50% of said exon-including nucleic acid probes comprise, contiguous to a first end of said fragment, a first intronic and/or intergenic sequence that is identically contiguous to said fragment in said eukaryotic genome, and further comprise, contiguous to a second end of said fragment, a second intronic and/or intergenic sequence that is identically contiguous to said fragment in said eukaryotic genome.

22. The nucleic acid microarray of claim 16, wherein at least 50% of said exon-including nucleic acid probes comprise at least a first priming sequence not found in contiguity with the rest of said probe sequence in said eukaryotic genome.

23. The nucleic acid microarray of claim 22, wherein said first priming sequence is identical among said primer-including exon-specific probes.

24. The nucleic acid microarray of claim 16, wherein at least 50% of probes disposed on said microarray lack prokaryotic and bacteriophage vector sequence.

25. The nucleic acid microarray of claim 16, wherein at least 50% of probes disposed on said microarray lack homopolymeric stretches of A or T.

26. The nucleic acid microarray of claim 16, wherein at least 50% of the genes having a first exon fragment present among the exon-specific probes on said

microarray have at least a second exon fragment present among said exon-specific probes.

27. The nucleic acid microarray of claim 16, wherein said genome is a human genome.

28. The nucleic acid microarray of claim 16, wherein at least 50% of the exons are detectable for each gene detectable by an exon-specific probe on said microarray.

29. The nucleic acid microarray of claim 16, wherein said plurality of addressably disposed nucleic acid probes are noncovalently bound to said substrate.

30. The nucleic acid microarray of claim 16, wherein each of said plurality of exon-specific probes is derived, directly or indirectly, by amplification from genomic DNA.

31. A visual display presenting eukaryotic genomic sequence annotated with information about a predetermined biologic function, comprising:

a first visual element, each point along the length of which first visual element maps linearly and uniquely to a nucleotide of said genomic sequence;

a second visual element, first and second boundaries of which second visual element map linearly to a first and second nucleotide of said genomic sequence, wherein said first and second nucleotides delimit a region of said genomic sequence predicted to have said predetermined function; and

a third visual element, first and second boundaries of which third visual element map linearly to a first and second nucleotide of said genomic sequence, wherein said first and second nucleotides delimit a region of said genomic sequence experimentally confirmed to have said predetermined function.

32. The visual display of claim 31, wherein said predetermined function is expression in mRNA.

33. The visual display of claim 32, wherein regions of genomic sequence not confirmed as expressed are omitted.

34. The visual display of claim 31, wherein said display is electronic.

35. An addressable set of exon-specific nucleic acid probes for detecting alternatively spliced mRNA variants, comprising:

a plurality of exon-specific nucleic acid probes, each of said plurality of probes separately and addressably isolatable from said plurality,

wherein each of said plurality of exon-specific probes includes a fragment of no more than one exon of a eukaryotic genome, said fragment selectively hybridizable at high stringency to an expressed gene,

wherein said probes collectively include specifically hybridizable fragments of a plurality of exons of at least one gene,

wherein said genome has, on average, at least one intron per gene, and

wherein said plurality of probes averages at least 100 bp in length.

36. An addressable set of exon-specific nucleic acid probes for detecting alternatively spliced mRNA variants, comprising:

a plurality of exon-specific nucleic acid probes, each of said plurality of probes separately and addressably amplifiable from said plurality,

wherein each of said plurality of exon-specific probes includes a fragment of no more than one exon of a eukaryotic genome, said fragment selectively hybridizable at high stringency to an expressed gene,

wherein said probes collectively include specifically hybridizable fragments of a plurality of [illegible] at least one gene,

wherein said genome has, on average, at least one intron per gene, and

wherein said plurality of probes averages at least 100 bp in length.

37. The addressable set of exon-specific nucleic acid probes of claim 35, wherein each of said probes is derived, directly or indirectly, by amplification from genomic DNA.

38. The addressable set of exon-specific probes of claim 36, wherein each of said probes is derived, directly or indirectly, by amplification from genomic DNA.

39. A method of associating exons that belong to a common gene, comprising:

ordering the relatedness in patterns of expression of a plurality of exons,

wherein the exons most closely related in patterns of expression belong to a common gene.

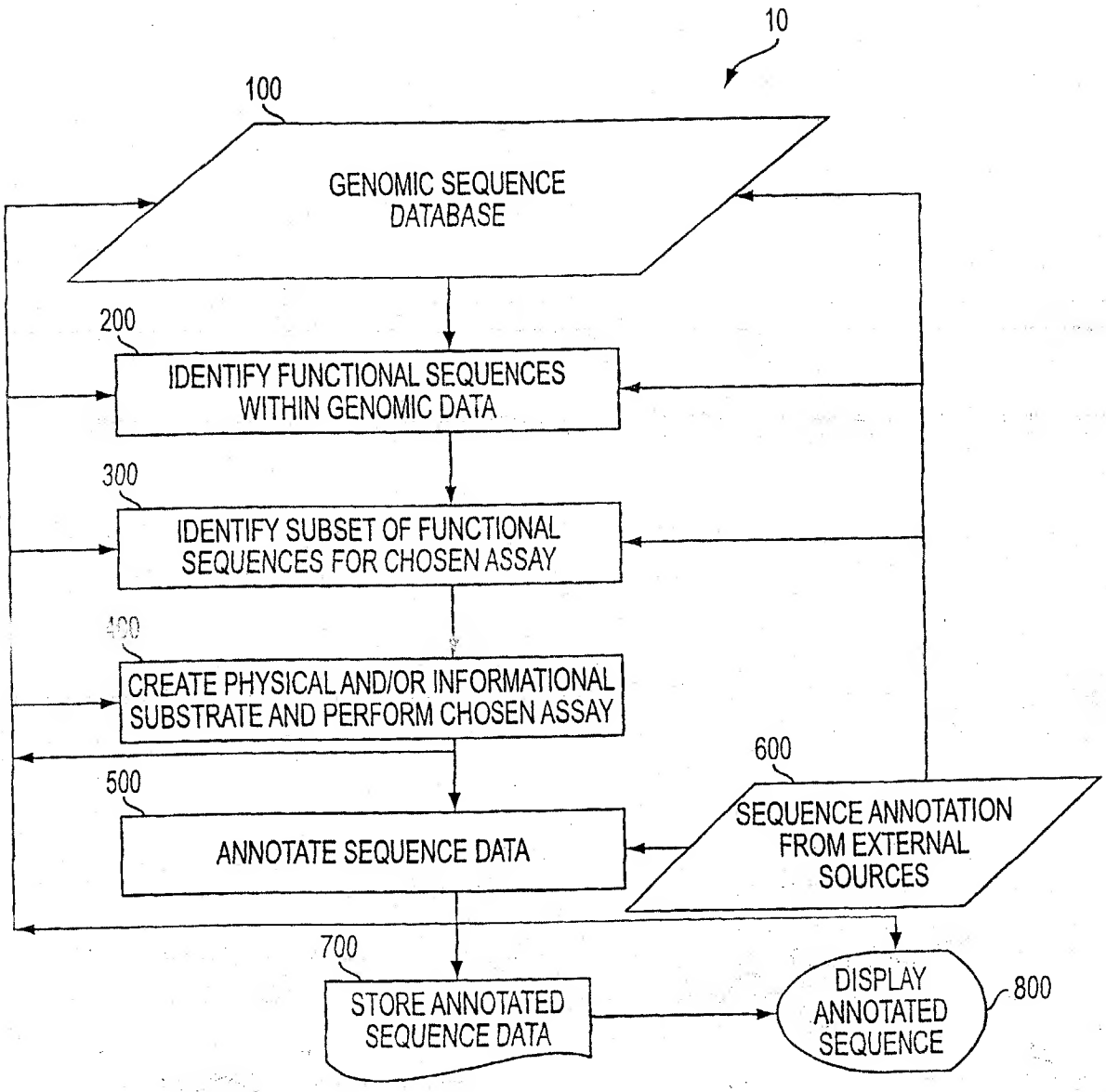

10
100
GENOMIC SEQUENCE DATABASE
200
IDENTIFY FUNCTIONAL SEQUENCES WITHIN GENOMIC DATA
300
IDENTIFY SUBSET OF FUNCTIONAL SEQUENCES FOR CHOSEN ASSAY
400
CREATE PHYSICAL AND/OR INFORMATIONAL SUBSTRATE AND PERFORM CHOSEN ASSAY
500
ANNOTATE SEQUENCE DATA
600
SEQUENCE ANNOTATION FROM EXTERNAL SOURCES
700
STORE ANNOTATED SEQUENCE DATA
800
DISPLAY ANNOTATED SEQUENCE

FIG. 1

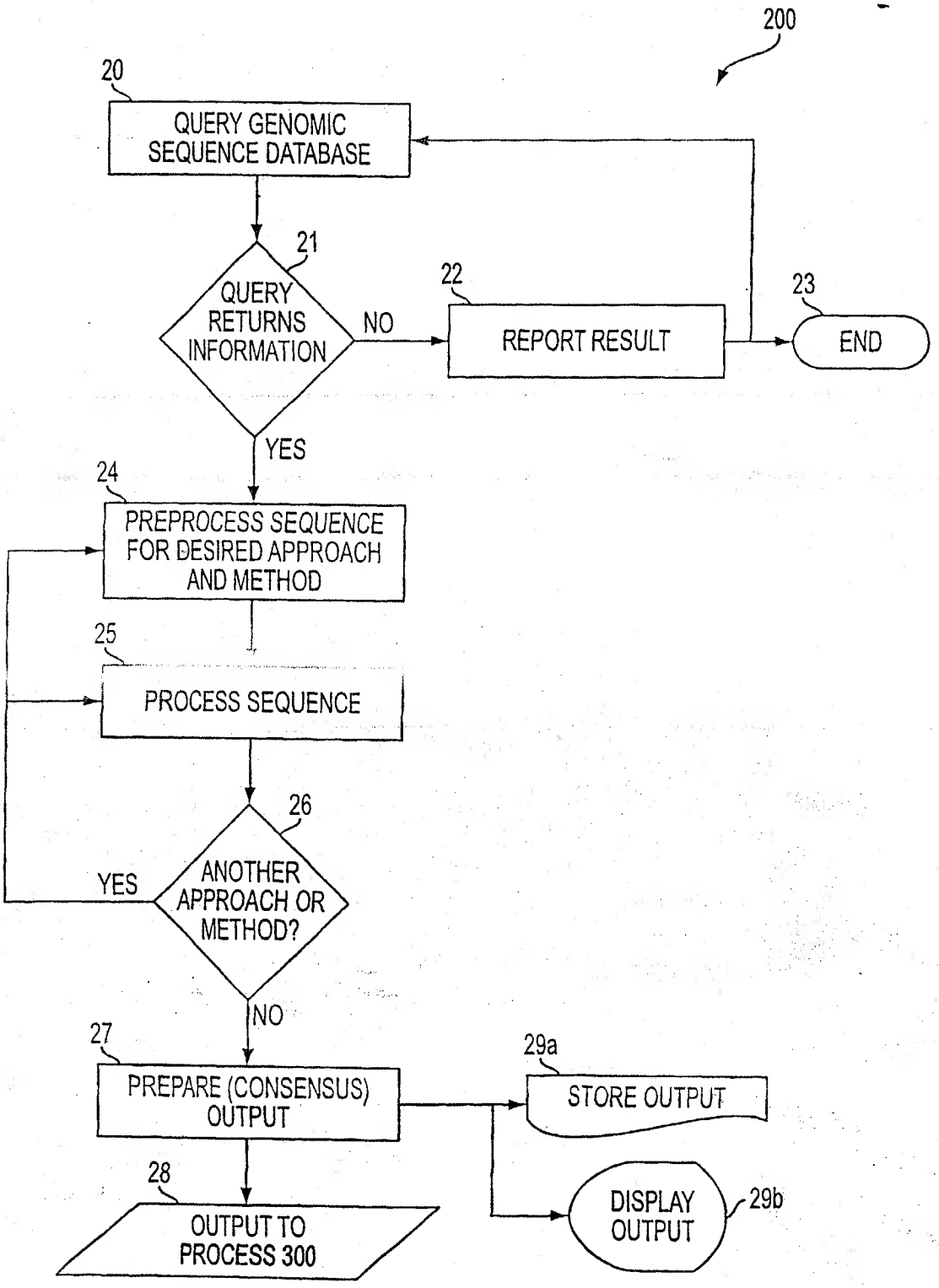

200
20
QUERY GENOMIC SEQUENCE DATABASE
21
QUERY RETURNS INFORMATION
NO
22
REPORT RESULT
23
END
YES
24
PREPROCESS SEQUENCE FOR DESIRED APPROACH AND METHOD
25
PROCESS SEQUENCE
26
ANOTHER APPROACH OR METHOD?
YES
NO
27
PREPARE (CONSENSUS) OUTPUT
29a
STORE OUTPUT
28
OUTPUT TO PROCESS 300
DISPLAY OUTPUT
29b

FIG. 2

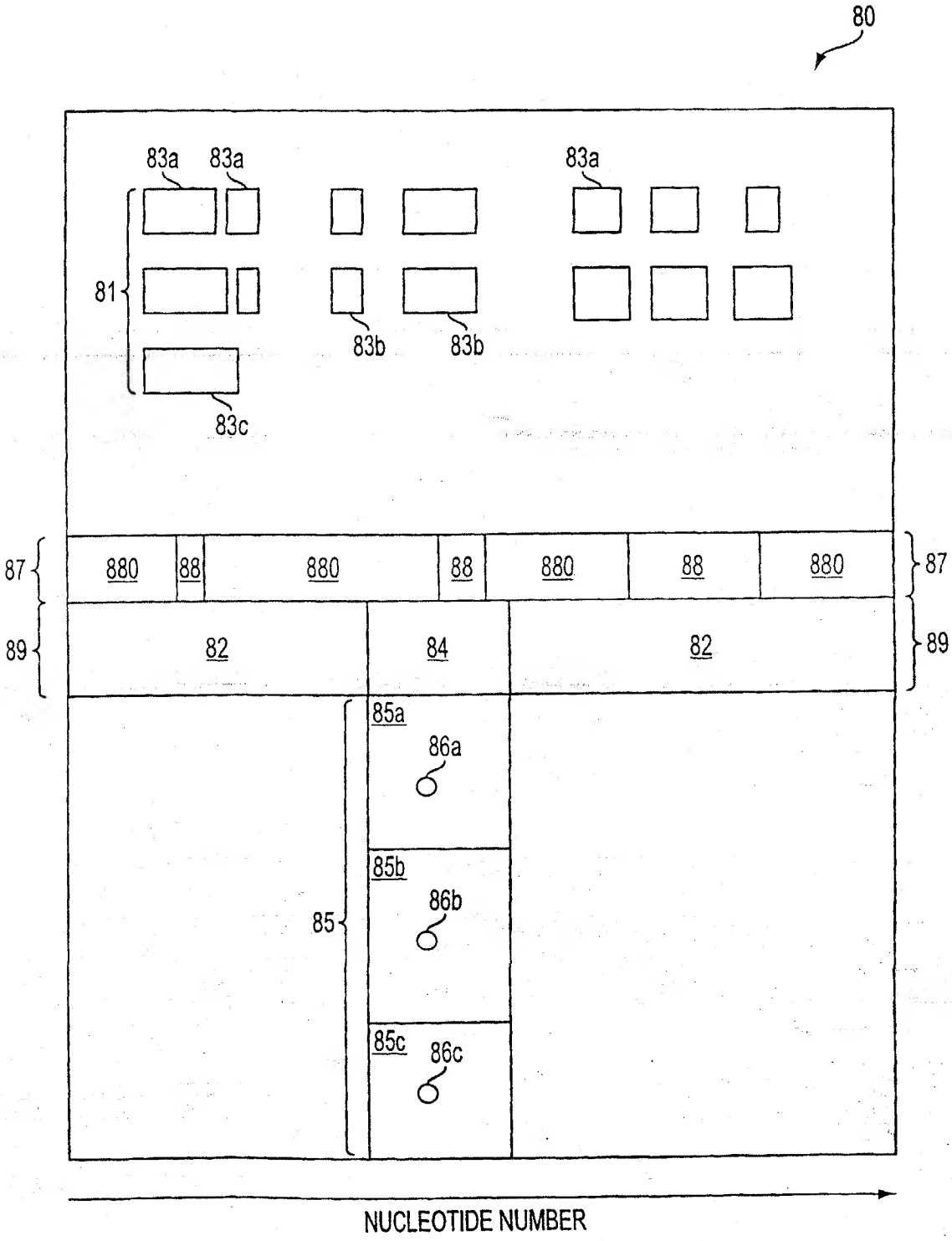

80
83a
83a
83a
81
83b
83b
83c
87
880
88
880
88
880
88
880
87
89
82
84
82
89
85a
86a
85b
86b
85
85c
86c
NUCLEOTIDE NUMBER

FIG. 3

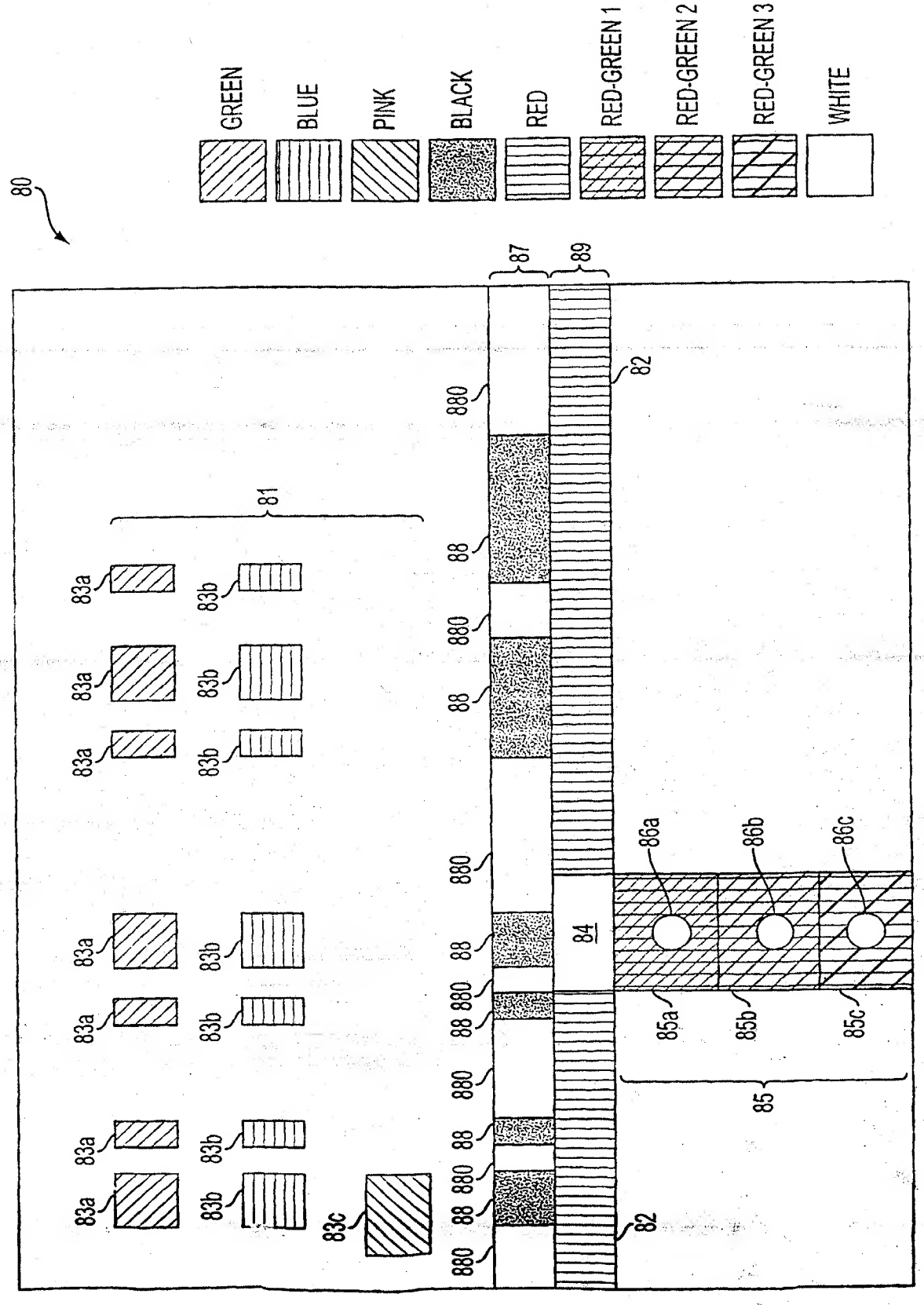

GREEN
BLUE
PINK
BLACK
RED
RED-GREEN 1
RED-GREEN 2
RED-GREEN 3
WHITE
80
81
82
83a
83b
83c
84
85
85a
85b
85c
86a
86b
86c
87
88
880
89

FIG. 4

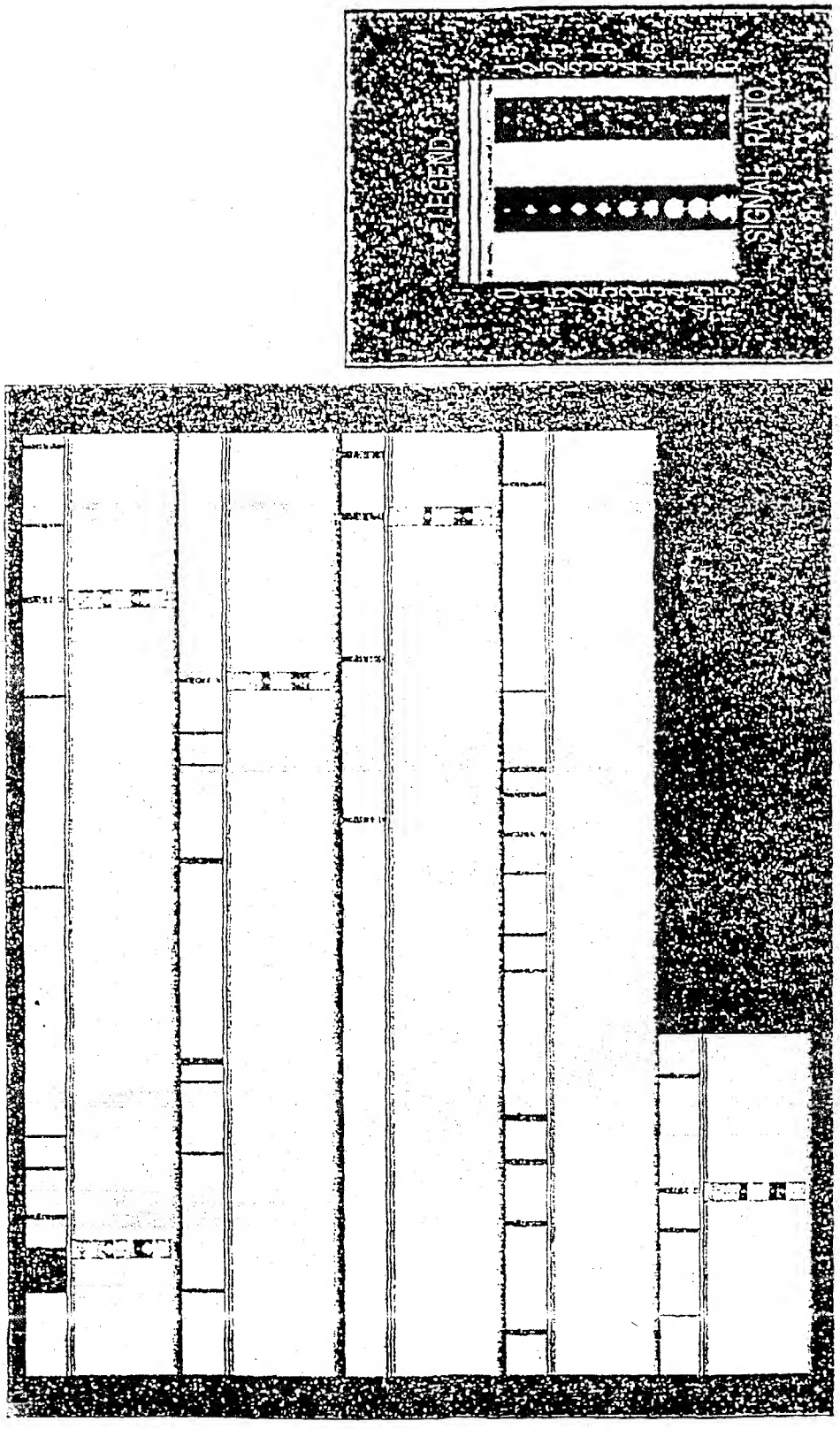


FIG. 5

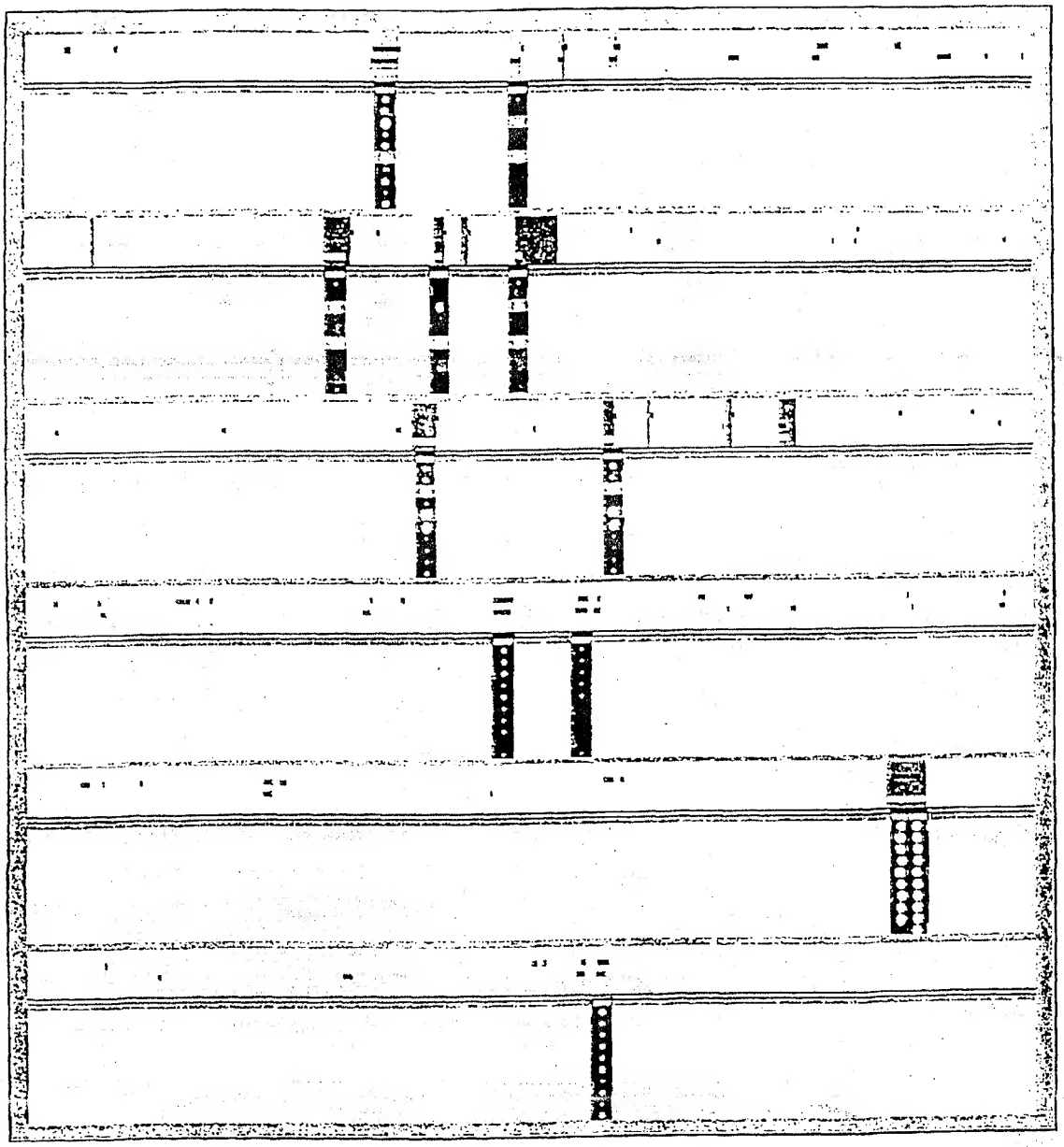


FIG. 6

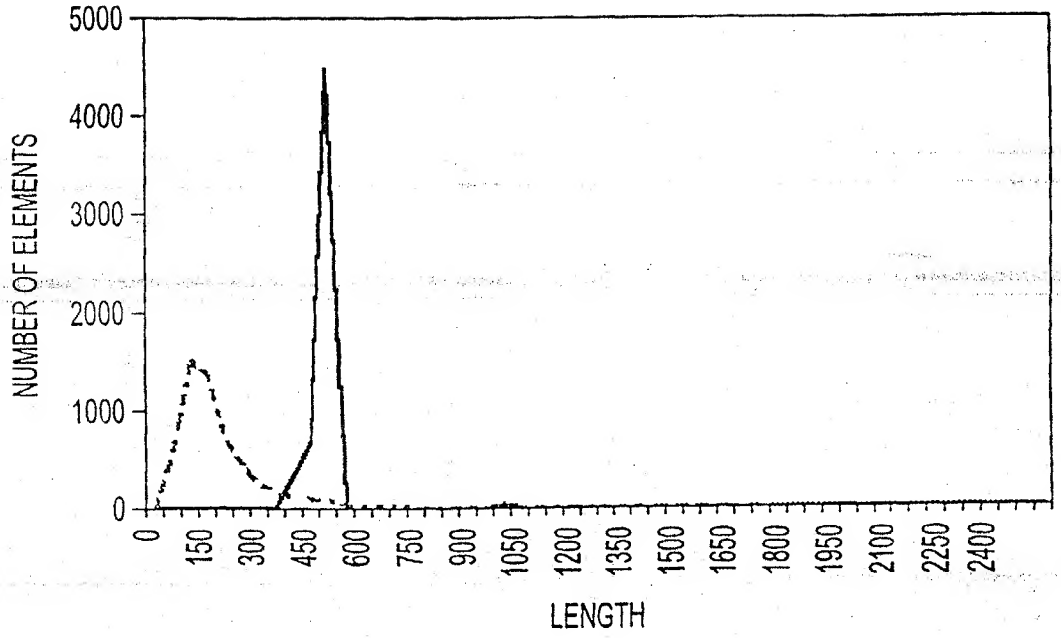

5000
4000
3000
2000
1000
0
NUMBER OF ELEMENTS
0
150
300
450
600
750
900
1050
1200
1350
1500
1650
1800
1950
2100
2250
2400
LENGTH

FIG. 7

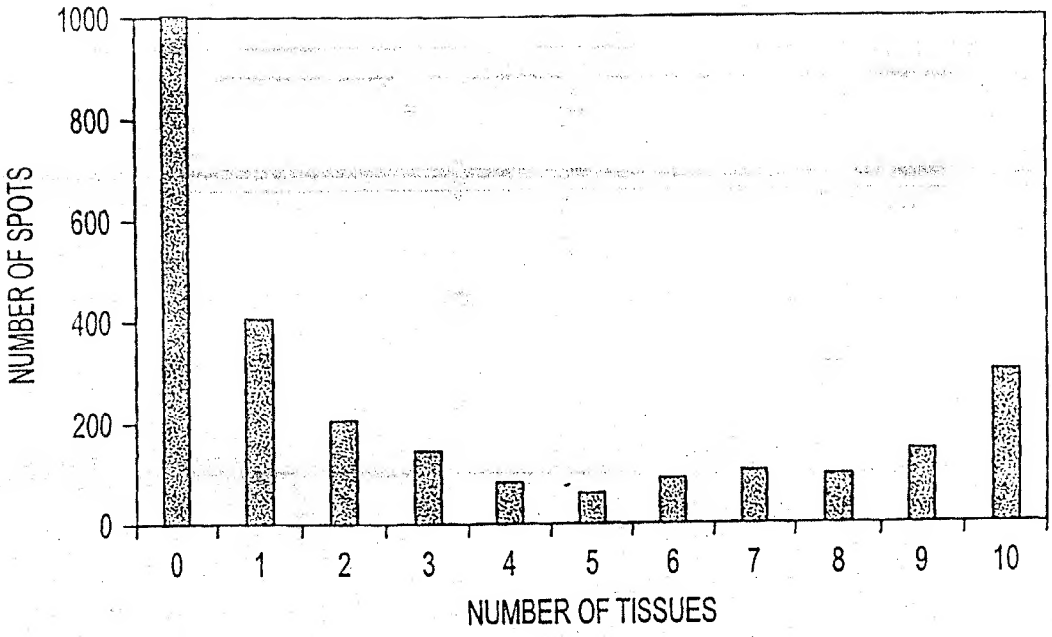

1000
800
600
400
200
0
NUMBER OF SPOTS
0
1
2
3
4
5
6
7
8
9
10
NUMBER OF TISSUES

FIG. 8

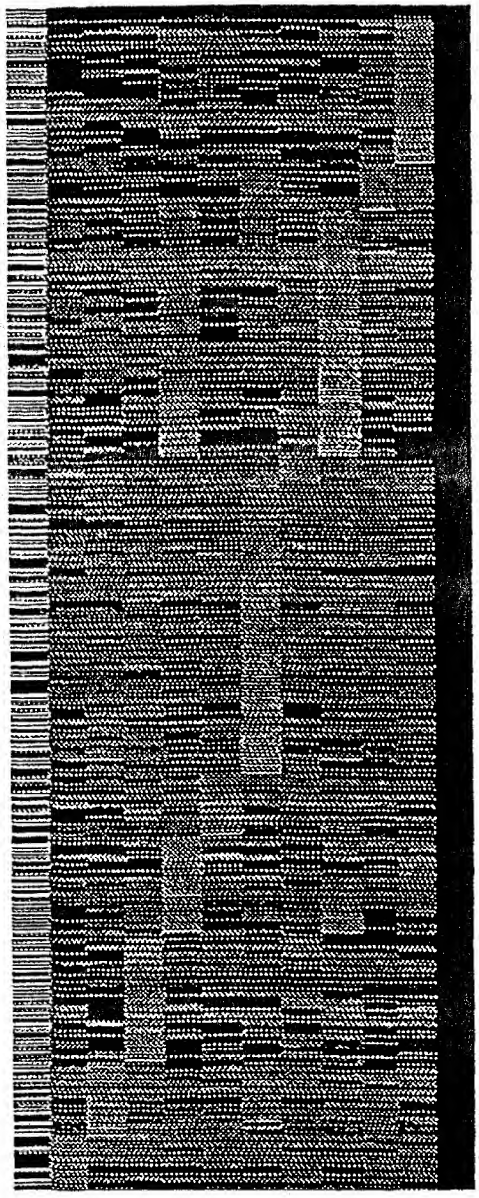


FIG. 9A

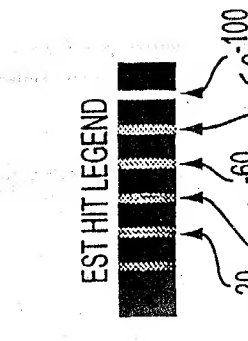


FIG. 9C

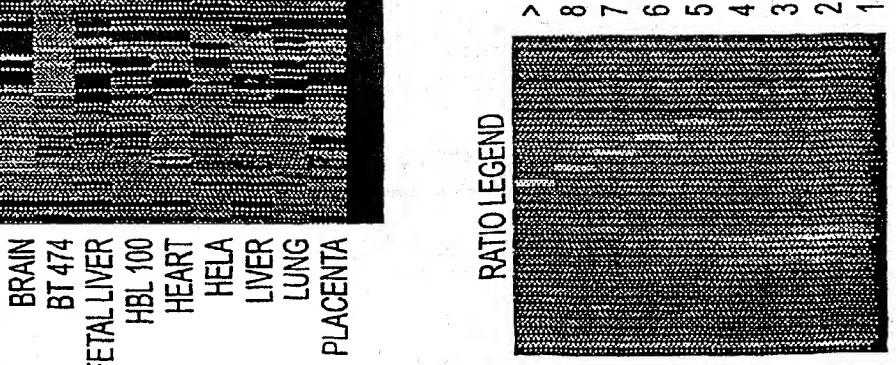


FIG. 9B

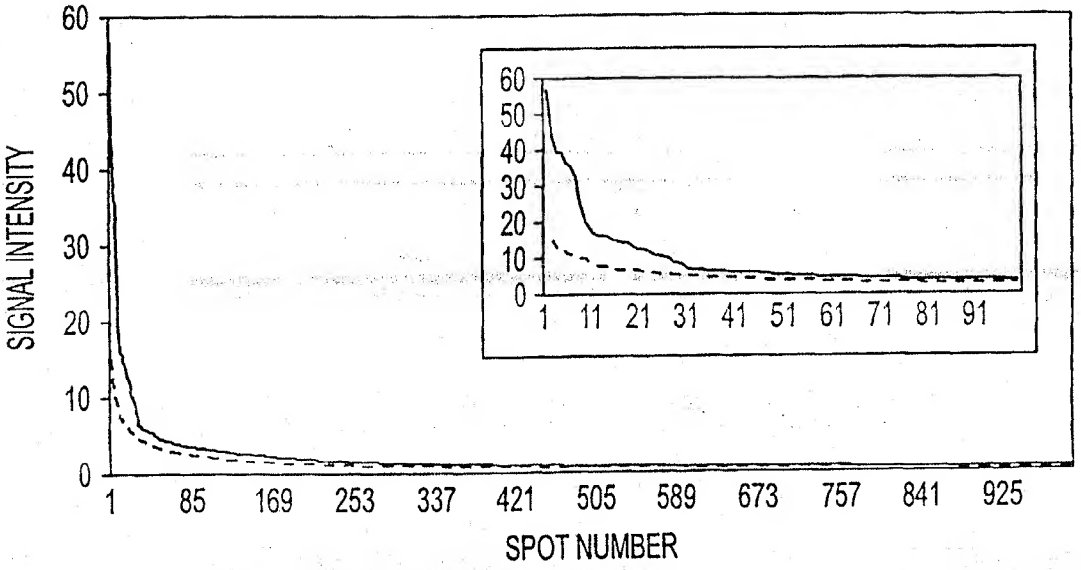

SIGNAL INTENSITY
60
50
40
30
20
10
0
1
85
169
253
337
421
505
589
673
757
841
925
SPOT NUMBER
60
50
40
30
20
10
0
1
11
21
31
41
51
61
71
81
91

FIG. 10

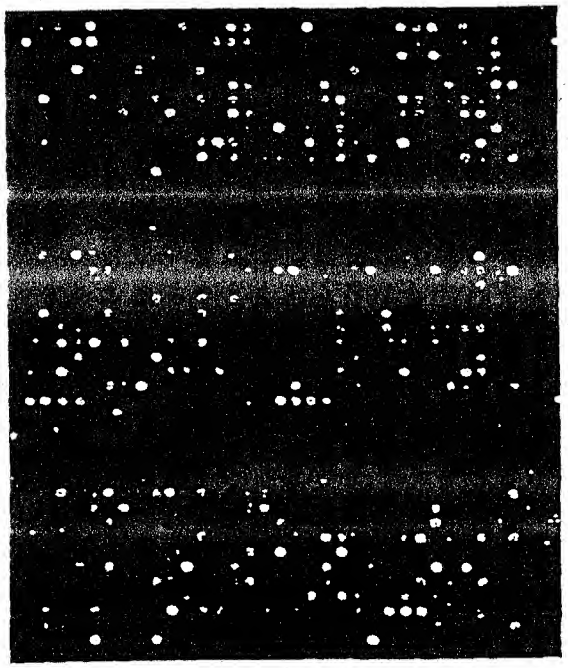


FIG. 11A

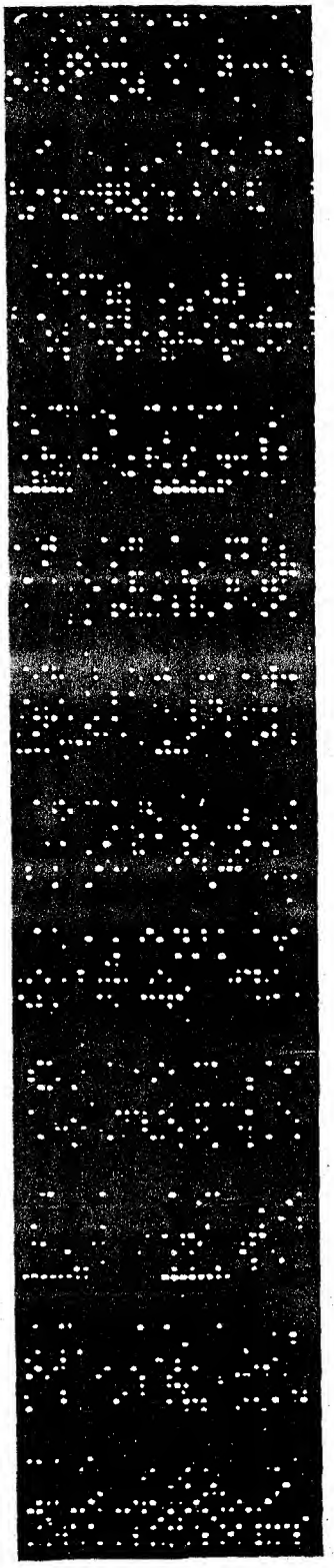


FIG. 11B

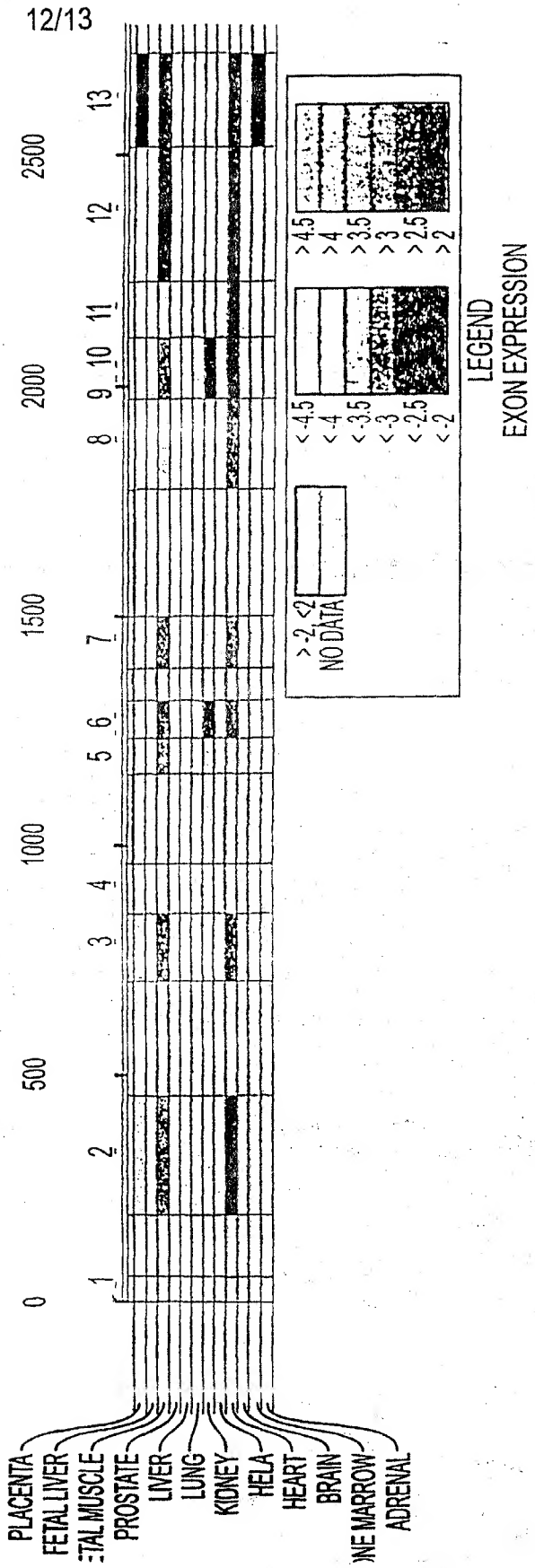

PLACENTA
FETAL LIVER
FETAL MUSCLE
PROSTATE
LIVER
LUNG
KIDNEY
HELA
HEART
BRAIN
BONE MARROW
ADRENAL
0
500
1000
1500
2000
2500
1
2
3
4
5
6
7
8
9
10
11
12
13
LEGEND
EXON EXPRESSION
> -2, <2
NO DATA
< -4.5
< -4
< -3.5
< -3
< -2.5
< -2
> 4.5
> 4
> 3.5
> 3
> 2.5
> 2

FIG. 12

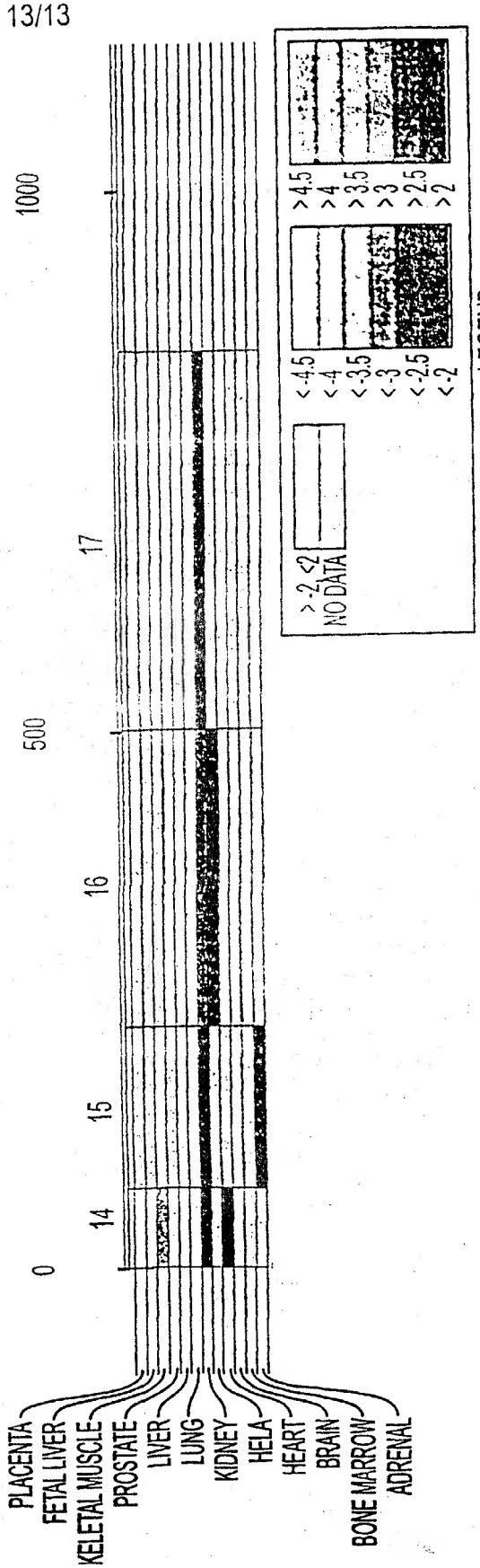

PLACENTA
FETAL LIVER
KELETAL MUSCLE
PROSTATE
LIVER
LUNG
KIDNEY
HELA
HEART
BRAIN
BONE MARROW
ADRENAL
0
500
1000
14
15
16
17
LEGEND
EXON EXPRESSION
> -2 <2
NO DATA
< -4.5
< -4
< -3.5
< -3
< -2.5
< -2
> 4.5
> 4
> 3.5
> 3
> 2.5
> 2

FIG. 13

SEQUENCE LISTING

<110> Molecular Dynamics, Inc.

<120> METHODS AND APPARATUS FOR HIGH-THROUGHPUT DETECTION AND CHARACTERIZATION OF ALTERNATIVELY SPLICED GENES

<130> MDhMORF-3 PCT

<150> GB 0024263.6
<151> 2000-10-04

<150> US 60/236,359
<151> 2000-09-27

<150> US 60/234,687
<151> 2000-09-21

<150> US 09/632,366
<151> 2000-08-03

<150> US 09/608,408
<151> 2000-06-30

<150> US 60/207,456
<151> 2000-05-26

<150> US 60/180,312
<151> 2000-02-04

<160> 17

<170> Molecular Dynamics Sequence Listing Engine

<210> 1
<211> 569
<212> DNA
<213> Homo sapiens

<220>
<223> MAP TO AL008582.11

<400> 1
attgcgttca cagggttctg gcgcctctag ggaggcctcg ccttcaccgt gacgccccac 60
tcttccgggc acgcccctgc ccaaaggctt taaaggcgcc gtgtgggacg tcactttaat 120
gcgacctcat ctttgtcagt gcacaaaatg gcgccctaca gcctactggt gactcggctg 180
caggtgagcg agctcaggga cctctgggtt cacgggggcg gggtgcctcc tactgtgccg 240
gcggctgtgg gcgaggcagg gcgaggcggg cccaacctgg ggccaacttc tcgtgtacct 300
gtcccggttg tgggcccggc acccatgccc gcttctctgg tgccctaggt caaggcgtcc 360
ccggcacagt cagctttcct ggccctgtcc ctgcctcgca agaagcgtgg gcccaagcag 420
cggcggccgg gtgaagagcg gaggcacctc tttcttcttt ttgaggacag cattgcctgc 480

tcccgtgggg tgtgtttcca ttcctcaagg tgaggacaag gtgactaggg ggcgcggtcg 540
ttgttgagtc ttctaggccc agttcctga 569

<210> 2
<211> 561
<212> DNA
<213> Homo sapiens

<220>
<223> MAP TO AL023553.5

<400> 2
gtgaacagga cccagcagtc cccacccaag tggaacgttg aggtcctgag ttcagactcc 60
ctgtccctgg ccactgttga ggttgccaca tggactgaga gggagaggca gggcgggagg 120
aggccgtgca gctagcacca ggccacccct ctgctcttct ccccacagac tgaaccggcc 180
gctgacactc tcggagaaga ttgtgtatgg acacctggat gaccccgcca gccaggaaat 240
tgagcgaggc aagtcgtacc tgcggctgcg gccggaccgt gtggccatgc aggatgcgac 300
ggcccagatg gccatgctcc agttcatcag cagcgggctg tccaaggtgg ctgtgccatc 360
caccatccac tgtgaccatc tgattgaagc ccaggttggg ggcgagaaag acctgcgccg 420
ggccaaggtg agcagaaggt ggctttgggg gtgggcaagt gggcaagact gggcgagagg 480
cctcaccctc acactggagc aaaccagggc attgcctcca aattctcaca cctctttgga 540
ttggtctgct tttaaaactt g 561

<210> 3
<211> 567
<212> DNA
<213> Homo sapiens

<220>
<223> MAP TO AL023553.5

<400> 3
agggactcat tctgggctgg ctgtggggtg gtggttggtg gggatgaacg ggagacggtg 60
ggacccagga gggaaaaggg aacaagttag actcgaatct tctgggaggg aggtagagac 120
caataagcag caatgtgaat ggcagcaggg ccatcctgac ttcgtggctg gcacaggcac 180
acacggcctc tcacagccgc ctcgcccccc cctgtccagg tgattggcgt gaagctgacg 240
ggctctctct ccggttggtc ctcacccaaa gatgtgatcc tgaaggtggc aggcatcctc 300
acggtgaaag gtggcacagg tgcaatcgtg gaataccacg ggcctggtgt agactccatc 360
tcctgcactg gtgaggaagg cggccaggcg acgtggcccc taccctgtgc tgggcctgat 420
gggtctccag ttgggagtag aagcggtgaa tggccttcac ttgagaatct gtctgttcca 480
tttggggact tgggatagac agaggtagaa ggaaaattgg agacagcatt agagattgtc 540
tagtccacat cctcattaaa cagatga 567

<210> 4
<211> 552
<212> DNA
<213> Homo sapiens

<220>
<223> MAP TO AL023553.5

<400> 4

cttcattcta cccagcagcc accctgagct ggggccctga gctctgatcc ccatggcctc 60
ccctcaccgg cccaggtttt tacccagggc ctcaagggat gagtcggccc gctgtcacca 120
catctctctg agtgggagga gaagcaatgc agctccctgg tgtgaagatg caggtggccg 180
cgtagcaggt gtgtgggacc ctggcccggc caccagccaa tgcccggggc tctgttgctc 240
cacaggcatg gcgacaatct gcaacatggg tgcagaaatt ggggccacca cttccgtgtt 300
cccttacaac cacaggatga agaagtacct gagcaagacc ggccgggaag gtgagctggc 360
aggggcaggc ccgtgtgggt ggaacagtca catgcccgct cctccccaag acagagctat 420
gcctttccct gtagggcagg ggttcttaac ccattgtctc cagacaggcg tccgaggctc 480
tgagaaagcc ctgaagaggt gcaatatgtt tctgagtaga ttgtctggag ctctctccag 540
cttctcaaag gg 552

<210> 5
<211> 594
<212> DNA
<213> Homo sapiens

<220>
<223> MAP TO AL023553.5

<400> 5
aatgtgttca gcatcatggc caatggctac agcatctttc ttcaaatgac gccacgtcat 60
gagttagcct gtcctttatt gttggatatt tcggttgctt gcaactggtc tctgttaaaa 120
aagtctctgg cccagcccag atggtgaact cttttggcca agtctggata atttgaagtc 180
aggtggagac ctctgctcac tgtctcctcc tgacccttaa ccccaccacc cacaatgcac 240
caggtctaat tggtagctgc accaattcaa gctatgaaga tatggggcgc tcagcagctg 300
tggccaagca ggcactggcc catggcctca agtgcaagtc ccagttcacc atcactccag 360
gttccgagca gatccgcgcc accattgagc gggacggcta tgtgagtgcc catatccccc 420
tgcccatctc ccccacccca tgctgagtaa tgcctccagg cggcacaagc ccagaggcct 480
gttgggccgg ggctggggca ggtcccagtg gctgccctgg ggttccagta cagcacttcg 540
tcctgcagcc accacatcac cccttcccat cagactctca cccacccttg acat 594

<210> 6
<211> 583
<212> DNA
<213> Homo sapiens

<220>
<223> MAP TO AL023553.5

<400> 6
agtgcaagtc ccagttcacc atcactccag gttccgagca gatccgcgcc accattgagc 60
gggacggcta tgtgagtgcc catatccccc tgcccatctc ccccacccca tgctgagtaa 120
tgcctccagg cggcacaagc ccagaggcct gttgggccgg ggctggggca ggtcccagtg 180
gctgccctgg ggttccagta cagcacttcg tcctgcagcc accacatcac cccttcccat 240
cagactctca cccacccttg acattctgtc ttcctctctc cctggcaggc acagatcttg 300
agggatctgg gtggcattgt cctggccaat gcttgtggcc cctgcattgg ccagtgggac 360
aggtaagagg cgtatctttt gacaagacag ccccttgtgc acagggtaca gagccccaga 420
agttggaggg ggaattattg gggtggagag aagagactcc agccaaggtc tctagctcca 480
gggactcttg cccattagaa gcctctggca gacatgcctg ggaagagggg ctgggtgagg 540
aagggccctg cagaggcact gggggggcaac ttggctgaag cct 583

<210> 7

<211> 566
<212> DNA
<213> Homo sapiens

<220>
<223> MAP TO AL023553.5

<400> 7
ctctggctta cgtgtggttg caggctgggc ccagtgatcc tgaggttccc ttctgctctg 60
cgcgtggccc caggaggagg caaccctggc agggcctctt cattttccct cggtaggagc 120
taggctgggc tgcgccacag gaacccagct tatctgtcct cgggacaggc caggtgacaa 180
ggccagatat ccctaaccct gatccctctg acctggcagg aaggacatca agaaggggga 240
gaagaacaca atcgtcacct cctacaacag gaacttcacg ggccgcaacg acgcaaaccc 300
cgagacccat gcctttgtca cgtccccaga ggtgagactg cccagctgcg cacaagcctg 360
ggatggcctc tgggggtccc tggcgggtca gaggaggagg cagaaggaga tggggactgg 420
ggtcatccaa gtggtagcca ggagctacag gccttcccag cctcaggcgc atgcttggtg 480
cttcctgcct ggggctccct gggtcatggg attatgagat atttatacag tggtttgtgc 540
ttatgagcat ggaatttgga atctca 566

<210> 8
<211> 570
<212> DNA
<213> Homo sapiens

<220>
<223> MAP TO AL023553.5

<400> 8
gagcgaacat tgacctgtcc caactttggg cggcctctgc cccataaggg agactgagca 60
gccagaggcc tttgagggga tgaaggcctg gcctgagccc atgtggcctt agggtggaag 120
caccaggacc acagaacacg tgtctgaaga cttgcctgcc tctcaccccc ctgtcacccc 180
tcctgggccc cggggcctgc tgcctgcctc tggagggctt gtcatccacc cctccagggc 240
catgccctga cctctgtcct ctctacttac cacccaaggt caaagggaag tgtaccactg 300
accacatctc agctgctggc ccctggctca agttccgtgg gcacttggat aacatctcca 360
acaacctgct cattggtgcc atcaacattg aaaacggcaa ggccaactcc gtgcgcaatg 420
ccgtcactca ggagtttggc cccgtccctg acactgcccg ctactacaag gtgggtcaga 480
gttgataggg gcaatgccag tggtcactcc tgaaggggcc tgcaaggcag gtgcagggag 540
gacattaggg gagtggaaac tgggaaggag 570

<210> 9
<211> 555
<212> DNA
<213> Homo sapiens

<220>
<223> MAP TO AL023553.5

<400> 9
cacagtgagc aggcagagag ggtctgaggt gattggactt tttctgcttt gagaaacaaa 60
cagaaccagg gctgaaccca agtcctggcc cagccgggtg aaaggactct ggcacccccc 120
ggtggctggg tggggcagag ggtgctccca ggaaggggc gccttgagct tcacagatgc 180
atcttgtgtg gggcccggag gccgtccctg tctcacccaa cctccctcca cacacacctg 240

cctctgccaa gcaccaatgg gtggcttctg tcttctttgc cactgcaaac aaccacgtgc 300
ctctgtcccc tcggggcctc gtttgggtct cattcacgca ggcttcactt gcccttaggc 360
agcaggcgag gaagggcccc tccagcccct ttaccgggag cctcaggatg cccaggcgcc 420
aggtgggtga ggccaggcag gtagggccag acaggtgagg acggtgccct cctctgcctt 480
ataaccttac ccccgcttgc ctgacagaaa catggcatca ggtgggtggt gatcggagac 540
gagaactacg gcgag 555

<210> 10
<211> 577
<212> DNA
<213> Homo sapiens

<220>
<223> MAP TO AL023553.5

<400> 10
tctttgccac tgcaaacaac cacgtgcctc tgtcccctcg gggcctcgtt tgggtctcat 60
tcacgcaggc ttcacttgcc cttaggcagc aggcgaggaa gggcccctcc agccccttta 120
ccgggagcct caggatgccc aggcgccagg tgggtgaggc caggcaggta gggccagaca 180
ggtgaggacg gtgccctcct ctgccttata accttacccc cgcttgcctg acagaaacat 240
ggcatcaggt gggtggtgat cggagacgag aactacggcg agggctcgag ccgggagcat 300
gcagctctgg agcctcgcca ccttgggggc cgggccatca tcaccaagag ctttgccagg 360
atccacggtg agctggagtc tgtacccagg ccatcctcat cccatcccta gtgatcaagg 420
tcactctccc tgcccgtggc tgagttgggc ctggttctag gctgtgtcca ctgcagccca 480
caggcccgtc agcctcttgc cccttcttag gctcacacag tgcacatccg acgctcagct 540
tcccggcttc ccgcaggccc tgcttccagg cttgtag 577

<210> 11
<211> 550
<212> DNA
<213> Homo sapiens

<220>
<223> MAP TO AL023553.5

<400> 11
cagcgcacac ttgctagggg cacccctagt gaaagggagc agaccagggc cccatagtca 60
ctgcccgggc attgtcccag gcagcaggat taggggcatc tcccagagcc ccagatgggt 120
tcagaaaatg aagctctcca ggctagtcag gcccccgatg accgaatgcc gcctgctttc 180
cagagaccaa cctgaagaaa cagggcctgc tgcctctgac cttcgctgac ccggctgact 240
acaacaagat tcaccctgtg gacaagctga ccattcaggg cctgaaggac ttcacccctg 300
gcaaggttag gggcccgggt ccccctgagg tggtggggtg aggggcagcc accttgtttc 360
ccctcctgca ctggccccag ggtagcttct cccaggaggc ttcattccag ctggaaaggc 420
ccccagttct ccaggtggcc ccacagagaa agcaaagtgg cttctcagag ttgggggttg 480
gagtcaaccc ggggccctca cacctcccca acctcccttt actcaccagg acctggcact 540
caggggacag 550

<210> 12
<211> 586
<212> DNA
<213> Homo sapiens

<220>
<223> MAP TO AL023553.5

<400> 12
cagagttggg ggttggagtc aacccggggc cctcacacct ccccaacctc cctttactca 60
ccaggacctg gcactcaggg gacagcccac ccactgcagg accctctggg ccccaggaat 120
cccctgtagg tgccacctgg gtctgacctg ggccatcagg cacagactgg cctaggattt 180
ggtttgcctg ctgacctctt aggtccccag gcagtgccct gtctccctga ccccccctgcg 240
gggccaaggg cacacagtac ccaccacttc cacccacacc caccttctcc ttgcagcccc 300
tgaagtgcat catcaagcac cccaacggga cccaggagac catcctcctg aaccacacct 360
tcaacgagac gcagattgag tggttccgcg ctggcagtgc cctcaacaga atgaaggaac 420
tgcaacagtg agggcagtgc ctccccgccc cgccgctggc gtcaagttca gctccacgtg 480
tgccatcagt ggatccgatc cgtccagcca tggcttccta ttccaagatg gtgtgaccag 540
acatgcttcc tgctcccccgc ttagcccacg gagtgactgt ggttgt 586

<210> 13
<211> 570
<212> DNA
<213> Homo sapiens

<220>
<223> MAP TO AL023553.5

<400> 13
acttttttagc ccccgtcttc ctattttgag tttggttcag atcttaagca gctccatgca 60
actgtattta tttttgatga caagactccc atctaaagtt tttctcctgc ctgatcattt 120
cattggtggc tgaaggattc tagagaacct tttgttcttg caaggaaaac aagaatccaa 180
aaccagtgac tgttctgtga gtgattggtg tctgtgccgt ttgttgtcaa gtccagggtc 240
caggaagggt cttttcagtc aaggtcagtg gaggcccaac agccacagcc acagatggat 300
ccatcactgc agtgaaggag gagcagggcc tagatggtgg aggagcgggg caagctggtg 360
ggcactcctg gccttgcctc actgtccagc tcgagacact atctccttag catcggcctc 420
agctgctgtt gtcttcacag ccggccatac caccttcccg caggctggta gggtccactg 480
tccagcccca gggctagttg ctggtccacc aggagagaag gcccaggcct ggctcactga 540
tggatccctg ccagggataa agaacacgca 570

<210> 14
<211> 585
<212> DNA
<213> Homo sapiens

<220>
<223> MAP TO AC006427.13

<400> 14
tgtaccctac aaaaatgtca atgtcattaa aaaaaagaga gagagaaaga aggccaaata 60
gacataatga ctacaagcaa tgcataatcc tgaattggat tctggattag ggagaaaaat 120
catagccaaa aaggcagcat tgggacaatt aacaacattt gaatgcagac tatatattgg 180
atatcagttt taataaatgt aaacttcctg agtttgataa ttgcattgtg attatattgg 240
aaaatgtcac tttctttagg gtagtgacat tgaaatattt ggagtgaaat gtaagatctc 300
cacaaatggt tcagcaaaag ataaaaataa gaatagtgaa aggaataaca atgtgagcac 360
atttgtgtgt agagaggaag atagagcaaa catgtggcac aaatgttaac agtaggtaaa 420
tctagatgaa gggtaggttc ttgcaacttt gctacagttg tttttacaaa ataaaaagtt 480

aattaaaaac aaaattaaag atttaaaagta agattttgaa tgggattatg caaatttgga 540
tgtcacttca agagaataaa cttgggcgat agattatttt atgca 585

<210> 15
<211> 576
<212> DNA
<213> Homo sapiens

<220>
<223> MAP TO AC006427.13

<400> 15
aaagattaaa agtaagattt tgaatgggat tatgcaaatt tggatgtcac ttcaagagaa 60
taaacttggg cgatagatta ttttatgcaa cagaaaacca ctggcaaatt taaaccataa 120
actgctctac cattaggata ctggattctt tagagctgaa ggaatattgg gaaaccagct 180
tcagaggctg tgcttgtagg gaaaatggca aaattcaact tccttgcggg aggacaacac 240
tgtgtcattg ctgagatgag gcatggcagg gtgcaccacc agtactgcct gatagatgca 300
gacagtgaaa ccatcaccat ggccactgct cctctagaag ctggatgctg ctgttgtagc 360
tgttgccacg tgctaccaga atgaattatc tgcttttctt cgcatcggta gctcacaatt 420
cacagcctaa ggcacctgta tctgattacc cgagtccaag gcacatacca aagcattggc 480
tgcatggaag ctaggaaagt cagaatctga atttcagggt ctgtaatgtc atcaagtctc 540
tgtccactgt catcaagtct cctaagggat gggctt 576

<210> 16
<211> 569
<212> DNA
<213> Homo sapiens

<220>
<223> MAP TO AC006427.13

<400> 16
cgaaaacaaa ctggttcatt cacagattct aatttttaa gagagatgag taggacacaa 60
tgagttgatc atcagtgttc agaatcgaca aagagaagcc agtcgcaggg caagttcccc 120
agccatccca tattttctct ctttttcttg gccttccttc tgcagccagg gaccttgaag 180
cccaaaggct gaaaagccta ttatcaccag atagaaatgt aagtaaacaa ggcatcacaa 240
taatgaatgc taaagcagaa ggcaagaatg gggggaagga cctcttcgct gcaaaggatt 300
gggctggaac tgggccattt tccaactacc ataataatag agggcttttg ttggtttcta 360
gacatgtaag tcaaaaagga tacttcaaga caagaatgga aggaggaaac ggtttcttca 420
ccattgaccc tcaatctgca ttccaaaatt ggcttcaagg ctaacaagct ctaatccttc 480
aaccaaacag agggagctaa aaagaatgga tttgaaaaga ccacaggaaa taagagtggc 540
tatagcactc gggacatagc cctagtcaa 569

<210> 17
<211> 561
<212> DNA
<213> Homo sapiens

<220>
<223> MAP TO AC006427.13

<400> 17

tttcagccca gtaagacctg ttttagactt ctaaccttca taattgtaag ataatcaatt 60
tttgttgttt taaatcataa agtttatggt aatttgttat aatagcaata gggagctaat 120
ttaaatttttg gtacctgtaa atgtggtgtt attgagataa atacctacaa atatagaatt 180
aactttagaa ttgtgcaatg ggaagaggct ggaagaattt tgaagtgcat gagagaagaa 240
aacctaaatt gcctcacaaa cactcttagc agagatattt ctgttaatga ctctgttagt 300
gaggacacag aagaaagtga ggatcatgat agggaaaact tacaacatct tagagaatgc 360
ttaaatcatt ataagcagat tgttggtgga aatatggaca ttaaaggcac tgccagtgag 420
ggcacagatg gaaataagga atatattatt gaaaaccaga ggaaaagaga tccttttttac 480
atagtgagag aaagtttagc tgaattgtgt acaggtatgt ggaaagcgaa tttgtaaaca 540
atgtgcttgg atatttagct g 561